

## (12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property  
Organization  
International Bureau



(43) International Publication Date  
11 November 2004 (11.11.2004)

PCT

(10) International Publication Number  
**WO 2004/096979 A2**

- (51) International Patent Classification<sup>7</sup>: **C12N** Harimon Street, 40 300 Kfar Yona (IL). **SOREK, Rotem** [IL/IL]; 15 Aharonovich Street, 76 564 Rehovot (IL).
- (21) International Application Number: **PCT/IL2004/000077** (74) Agent: **G.E. EHRLICH (1995) LTD.**; 11 Menachem Begin Street, 52 521 Ramat-Gan (IL).
- (22) International Filing Date: 27 January 2004 (27.01.2004) (81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data: 10/426,002 30 April 2003 (30.04.2003) US (71) Applicant (for all designated States except US): **COMPU-GEN LTD.** [IL/IL]; 72 Pinchas Rosen Street, 69 512 Tel Aviv (IL).
- (72) Inventors; and
- (75) Inventors/Applicants (for US only): **MINTZ, Liat** [US/US]; 2 Kulesa Court, East Brunswick, NJ 08816 (US). **XIE, Hanqing** [US/US]; 1221 Ravens Crest Drive, Plainsboro, NJ 08536 (US). **DAHARY, Dvir** [IL/IL]; 23 Pinkas Street, 62 662 Tel Aviv (IL). **LEVANON, Erez** [IL/IL]; 73 Menachem Begin Street, 49732 Petach Tikva (IL). **FREILICH, Shiri** [IL/IL]; 119 Aba Hushi Street, 34987 Haifa (IL). **BECK, Nili** [IL/IL]; 4/10 Mota Gur Street, 44405 Kfar Saba (IL). **ZHU, Wei-Yong** [US/US]; 308 Quail Ridge Drive, Plainsboro, NJ 08536 (US). **WASSERMAN, Alon** [US/US]; 564 First Avenue, Apt. 13a, New York, NY 10016 (US). **CHERMESH, Chen** [IL/IL]; 58b Kordova Street, 50297 Mishmar HaShiva (IL). **AZAR, Idit** [IL/IL]; 29 Shalom Aleichem Street, 63344 Tel Aviv (IL). **BERNSTEIN, Jeanne** [IL/IL]; 23
- (84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).
- Published:**
- without international search report and to be republished upon receipt of that report
  - with sequence listing part of description published separately in electronic form and available upon request from the International Bureau
- For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: METHODS AND SYSTEMS FOR ANNOTATING BIOMOLECULAR SEQUENCES

(57) Abstract: A method of annotating biomolecular sequences. The method comprises (a) computationally clustering the biomolecular sequences according to a progressive homology range, to thereby generate a plurality of clusters each being of a predetermined homology of the homology range; and (b) assigning at least one ontology to each cluster of the plurality of clusters, the at least one ontology being: (i) derived from an annotation preassociated with at least one biomolecular sequence of each cluster; and/or (ii) generated from analysis of the at least one biomolecular sequence of each cluster thereby annotating biomolecular sequences.

BEST AVAILABLE COPY

WO 2004/096979 A2

## METHODS AND SYSTEMS FOR ANNOTATING BIOMOLECULAR SEQUENCES

### FIELD AND BACKGROUND OF THE INVENTION

5       The present invention relates to systems and methods useful for annotating biomolecular sequences. More particularly, the present invention relates to computational approaches, which enable systemic characterization of biomolecular sequences and identification of differentially expressed biomolecular sequences such as sequences associated with a pathology.

10       In the postgenomic era, data analysis rather than data collection presents the biggest challenge to biologists. Efforts to ascribe biological meaning to genomic data, whether by identification of function, structure or expression pattern are lagging behind sequencing efforts [Boguski MS (1999) Science 286:453-455].

15       It is well recognized that elucidation of spatial and temporal patterns of gene expression in healthy and diseased states may contribute immensely to further understanding of disease mechanisms.

      Therefore, any observational method that can rapidly, accurately and economically observe and measure the pattern of expression of selected  
20 individual genes or of whole genomes is of great value to scientists.

      In recent years, a variety of techniques have been developed to analyze differential gene expression. However, current observation and measurement methods are inaccurate, time consuming, labor intensive or expensive, oftentimes requiring complex molecular and biochemical analysis of numerous  
25 gene sequences.

      For example, observation methods for individual mRNA or cDNA molecules such as Northern blot analysis, RNase protection, or selective hybridization to arrayed cDNA libraries [see Sambrook et al. (1989) Molecular cloning, A laboratory manual, Cold Spring Harbor press, NY] depend on  
30 specific hybridization of a single oligonucleotide probe complementary to the

known sequence of an individual molecule. Since a single human cell is estimated to express 10,000-30,000 genes [Liang et al. (1992) Science 257:967-971], single probe methods to identify all sequences in a complex sample are ineffective and laborious.

5 Other approaches for high throughput analysis of differential gene expression are summarized infra.

*EST sequencing* – The basic idea is to create cDNA libraries from tissues of interest, pick clones randomly from these libraries and then perform a single sequencing reaction from a large number of clones. Each sequencing  
10 reaction generates 300 base pairs or so of sequence that represents a unique sequence tag for a particular transcript. An EST sequencing project is technically simple to execute since it requires only a cDNA library, automated DNA sequencing capabilities and standard bioinformatics protocols.

To generate meaningful amounts of data, however, high throughput  
15 template preparation, sequencing and analysis protocols must be applied. As such, the number of new genes identified as well as the statistical significance of the data is proportional to the number of clones sequenced as well as the complexity of the tissue being analyzed [Adams et al. (1995) Nature 377:3-173; Hillier et al. (1996) Genome Res. 6:807-828].

20 *Subtractive cloning* –Subtractive cloning offers an inexpensive and flexible alternative to EST sequencing and cDNA array hybridization. In this approach, double-stranded cDNA is created from the two-cell or tissue populations of interest, linkers are ligated to the ends of the cDNA fragments and the cDNA pools are then amplified by PCR. The cDNA pool from which  
25 unique clones are desired is designated the “tester”, and the cDNA pool that is used to subtract away shared sequences is designated the “driver”. Following initial PCR amplification, the linkers are removed from both cDNA pools and unique linkers are ligated to the tester sample. The tester is then hybridized to a vast excess of driver DNA and sequences that are unique to the tester cDNA  
30 pool are amplified by PCR.

The primary limitation of subtractive methods is that they are not always comprehensive. The cDNAs identified are typically those, which differ significantly in expression level between cell-populations and subtle quantitative differences are often missed. In addition each experiment is a pair wise comparison and since subtractions are based on a series of sensitive biochemical reactions it is difficult to directly compare a series of RNA samples.

*Differential display* – Differential display is another PCR-based differential cloning method [Liang and Pardee (1992) Science 257:967-70; Welsh et al. (1992) Nucleic Acids Res. 20:4965-70]. In classical differential display, reverse transcription is primed with either oligo-dT or an arbitrary primer. Thereafter an arbitrary primer is used in conjunction with the reverse transcription primer to amplify cDNA fragments and the cDNA fragments are separated on a polyacrylamide gel. Differences in gene expression are visualized by the presence or absence of bands on the gel and quantitative differences in gene expression are identified by differences in the intensity of bands. Adaptation of differential display methods for fluorescent DNA sequencing machines has enhanced the ability to quantify differences in gene expression [Kato (1995) Nucleic Acids Res. 18:3685-90].

A limitation of the classical differential display approach is that false positive results are often generated during PCR or in the process of cloning the differentially expressed PCR products. Although a variety of methods have been developed to discriminate true from false positives, these typically rely on the availability of relatively large amounts of RNA.

*Serial analysis of gene expression (SAGE)* – this DNA sequence based method is essentially an accelerated version of EST sequencing [Valculescu et al. (1995) Science 270:484-8]. In this method a digestible unique sequence tag of 13 or more bases is generated for each transcript in the cell or tissue of interest, thereby generating a SAGE library.



Sequencing each SAGE library creates transcript profiles. Since each sequencing reaction yields information for twenty or more genes, it is possible to generate data points for tens of thousands of transcripts in modest sequencing efforts. The relative abundance of each gene is determined by counting or clustering sequence tags. The advantages of SAGE over many other methods include the high throughput that can be achieved and the ability to accumulate and compare SAGE tag data from a variety of samples, however the technical difficulties concerning the generation of good SAGE libraries and data analysis are significant.

Altogether, it is clear from the above that laboratory bench approaches are ineffective, time consuming, expensive and often times inaccurate in handling and processing the vast amount of genomic information which is now available.

It is appreciated, that much of the analysis can be effected by developing computational algorithms, which can be applied to mining data from existing databases, thereby retrieving and integrating valuable biological information.

To date, there are more than a hundred major biomolecule databases and application servers on the Internet and new sites are being introduced at an ever-increasing rates [Ashburner and Goodman (1997) *Curr. Opin. Genet. Dev.* 7:750-756; Karp (1998) *Trends Biochem. Sci.* 23:114-116].

However, these databases are organized in extremely heterogeneous formats. These reflect the inherent complexity of biological data, ranging from plain-text nucleic acid and protein sequences, through the three dimensional structures of therapeutic drugs and macromolecules and high resolution images of cells and tissues, to microarray-chip outputs. Moreover data structures are constantly evolving to reflect new research and technology development.

The heterogeneous and dynamic nature of these biological databases present major obstacles in mining data relevant to specific biological queries. Clearly, simple retrieval of data is not sufficient for data mining; efficient data retrieval requires flexible data manipulation and sophisticated data integration.

Efficient data retrieval requires the use of complex queries across multiple heterogeneous data sources; data warehousing by merging data derived from multiple public sources and local (i.e., private) sources; and multiple data-analysis procedures that require feeding subsets of data derived from different sources into various application programs for gene finding, protein-structure prediction, functional domain or motif identification, phylogenetic tree construction, graphic presentation and so forth.

Current biological data retrieval systems are not fully up to the demand of smooth and flexible data integration [Etzold et al. (1996) Methods Enzymol. 266:114-128; Schuler et al. (1996) Methods Enzymol. 266:141-162; Chung and Wong (1999) Trends Biotech. 17:351-355].

There is thus a widely recognized need for, and it would be highly advantageous to have, systems and methods which can be used for efficient retrieval and processing of data from biological databases thereby enabling annotation of previously un-annotated biomolecular sequences.

### SUMMARY OF THE INVENTION

According to one aspect of the present invention there is provided a method of annotating biomolecular sequences according to a hierarchy of interest, the method comprising: (a) computationally constructing a dendrogram having multiple nodes, the dendrogram representing the hierarchy of interest, wherein each node of the multiple nodes of the dendrogram is annotated by at least one keyword; (b) computationally assigning each biomolecular sequence of the biomolecular sequences to a specific node of the multiple nodes of the dendrogram to thereby generate assigned biomolecular sequences; and (c) computationally classifying each of the assigned biomolecular sequences to nodes hierarchically higher than the specific node, thereby annotating biomolecular sequences according to the hierarchy of interest.

According to another aspect of the present invention there is provided a method of identifying differentially expressed biomolecular sequences, the

method comprising: (a) computationally constructing a dendrogram having multiple nodes, the dendrogram representing the hierarchy of interest, wherein each node of the multiple nodes of the dendrogram is annotated by at least one keyword; (b) computationally assigning each biomolecular sequence of the biomolecular sequences to a specific node of the multiple nodes of the dendrogram to thereby generate assigned biomolecular sequences; (c) computationally classifying each of the assigned biomolecular sequences to nodes hierarchically higher than the specific node, to thereby generate annotated biomolecular sequences; and (d) identifying annotated biomolecular sequences assigned to a portion of the multiple nodes, thereby identifying differentially expressed biomolecular sequences.

According to yet another aspect of the present invention there is provided a computer readable storage medium comprising a database stored in a retrievable manner, the database including files each containing data of a specific node of a dendrogram, the data including biomolecular sequence information and biomolecular sequence annotations, wherein the biomolecular sequence annotations are selected from the group consisting of contig description, tissue specific expression, pathological specific expression, functional features, parameters for ontological annotation assignment, cellular localization, database sequence source and functional alterations.

According to still another aspect of the present invention there is provided a system for generating a database of annotated biomolecular sequences, the system comprising a processing unit, the processing unit executing a software application configured for: (a) constructing a dendrogram having multiple nodes, the dendrogram representing a hierarchy of interest, wherein each node of the multiple nodes of the dendrogram is annotated by at least one keyword; (b) assigning each biomolecular sequence of the biomolecular sequences to a specific node of the multiple nodes of the dendrogram to thereby generate assigned biomolecular sequences; (c) classifying each of the assigned biomolecular sequences to nodes hierarchically

higher than the specific node, to thereby generate annotated biomolecular sequences; and (d) storing sequence annotations and sequence information of the annotated biomolecular sequences, thereby generating the database of annotated biomolecular sequences.

5 According to further features in preferred embodiments of the invention described below, the biomolecular sequences are selected from the group consisting of polypeptide sequences and polynucleotide sequences.

According to still further features in the described preferred  
embodiments the polynucleotides are selected from the group consisting of  
10 genomic sequences, expressed sequence tags, contigs, complementary DNA  
(cDNA) sequences, pre-messenger RNA (mRNA) sequences, and mRNA  
sequences.

According to still further features in the described preferred  
embodiments the biomolecular sequences are selected from the group consisting  
15 of annotated biomolecular sequences, unannotated biomolecular sequences and  
partially annotated biomolecular sequences.

According to still further features in the described preferred  
embodiments the method further comprising homology clustering of the  
biomolecular sequences prior to step (b).

20 According to still further features in the described preferred  
embodiments the dendrogram is selected from the group consisting of a graph, a  
list, a map and a matrix.

According to still further features in the described preferred  
embodiments the hierarchy of interest is selected from the group consisting of a  
25 tissue expression hierarchy, a developmental expression hierarchy, a  
pathological expression hierarchy, a cellular expression hierarchy, an  
intracellular expression hierarchy, a taxonomical hierarchy and a functional  
hierarchy.

According to still further features in the described preferred  
30 embodiments each node of the multiple nodes is a parental node in an additional

hierarchy of interest.

According to still further features in the described preferred embodiments the method further comprising classifying the biomolecular sequences of the parental node according to the additional hierarchy of interest.

5 According to still further features in the described preferred embodiments the system further comprising classifying the biomolecular sequences of the parental node according to the additional hierarchy of interest.

According to still further features in the described preferred embodiments each of the biomolecular sequences is a member of a sequence  
10 contig.

According to still further features in the described preferred embodiments the method further comprising the step of confirming annotations of the assigned biomolecular sequence in-vivo and/or in-vitro prior to or following step (c).

15 According to still further features in the described preferred embodiments the system further comprising the step of confirming annotations of the assigned biomolecular sequence in-vivo and/or in-vitro prior to or following step (c).

According to an additional aspect of the present invention there is provided a method of identifying sequence features unique to differentially  
20 expressed mRNA splice variants, the method comprising: (a) computationally identifying unique sequence features in each splice variant of an alternatively spliced expressed sequences; and (b) identifying differentially expressed splice variants of the alternatively spliced expressed sequences, thereby identifying sequence features unique to differentially expressed mRNA splice variants.

25 According to yet an additional aspect of the present invention there is provided a computer readable storage medium comprising data stored in a retrievable manner, the data including sequence information of sequence features unique to differentially expressed mRNA splice variants as set forth in files:

30 "Transcripts\_nucleotide\_seqs\_part1",

"Transcripts\_nucleotide\_seqs\_part2"

"Transcripts\_nucleotide\_seqs\_part3.new"

"Transcripts\_nucleotide\_seqs\_part4" and/or

"Protein.seqs"

5 provided in CD-ROMs 1 and/or 2 enclosed herewith, and sequence annotations as set forth in annotation categories "#TAA\_CD" and "#TAA\_TIS", in the file "Summary\_table.new" of CD-ROM3 enclosed herewith.

According to still an additional aspect of the present invention there is provided a system for generating a database of sequence features unique to  
10 differentially expressed mRNA splice variants, the system comprising a processing unit, the processing unit executing a software application configured for: (a) identifying unique sequence features in each splice variant of an alternatively spliced expressed sequences; and (b) identifying differentially expressed splice variants of the alternatively spliced expressed sequences,  
15 thereby identifying sequence features unique to differentially expressed mRNA splice variants. (c) storing the sequence features unique to the differentially expressed mRNA splice variants, thereby generating the database of sequence features unique to differentially expressed mRNA splice variants.

According to still further features in the described preferred  
20 embodiments step (b) is effected by qualifying annotations associated with the alternatively spliced expressed sequences.

According to still further features in the described preferred embodiments the method further comprising scoring the annotations associated with the alternatively spliced expressed sequences according to: (i) prevalence of the  
25 alternatively spliced expressed sequences in normal tissues; (ii) prevalence of the alternatively spliced expressed sequences in pathological tissues; (iii) prevalence of the alternatively spliced expressed sequence in total tissues; and (iv) number of tissues and/or tissue types expressing the alternatively spliced expressed sequences;

According to still further features in the described preferred embodiments the system further comprising scoring the annotations associated with the alternatively spliced expressed sequences according to: (i) prevalence of the alternatively spliced expressed sequences in normal tissues; (ii) prevalence of the alternatively spliced expressed sequences in pathological tissues; (iii) prevalence of the alternatively spliced expressed sequence in total tissues; and (iv) number of tissues and/or tissue types expressing the alternatively spliced expressed sequences;

According to still further features in the described preferred embodiments step (b) is effected by identifying the unique sequence feature.

According to still further features in the described preferred embodiments the unique sequence feature is selected from the group consisting of a donor-acceptor concatenation, an alternative exon, an exon and a retained intron.

According to still further features in the described preferred embodiments identifying unique sequence features in each splice variant of an alternatively spliced expressed sequence is effected by expressed sequence alignment.

According to a further aspect of the present invention there is provided a kit useful for detecting differentially expressed polynucleotide sequences, the kit comprising at least one oligonucleotide being designed and configured to be specifically hybridizable with a polynucleotide sequence selected from the group consisting of sequence files:

"Transcripts\_nucleotide\_seqs\_part1"

"Transcripts\_nucleotide\_seqs\_part2"

"Transcripts\_nucleotide\_seqs\_part3.new" and

"Transcripts\_nucleotide\_seqs\_part4",

provided in CD-ROMs 1 and/or 2 enclosed herewith, under moderate to stringent hybridization conditions.

According to still further features in the described preferred embodiments the at least one oligonucleotide is labeled.

According to still further features in the described preferred embodiments

the at least one oligonucleotide is attached to a solid substrate.

According to still further features in the described preferred embodiments the solid substrate is configured as a microarray and whereas the at least one oligonucleotide includes a plurality of oligonucleotides each being capable of  
5 hybridizing with a specific polynucleotide sequence of the polynucleotide sequences set forth in the files:

“Transcripts\_nucleotide\_seqs\_part1”

“Transcripts\_nucleotide\_seqs\_part2”

“Transcripts\_nucleotide\_seqs\_part3.new”and/or

10 “Transcripts\_nucleotide\_seqs\_part4”,

provided in CD-ROMs 1 and/or 2 enclosed herewith.

According to still further features in the described preferred embodiments each of the plurality of oligonucleotides is being attached to the microarray in a regio-specific manner.

15 According to still further features in the described preferred embodiments the at least one oligonucleotide is designed and configured for DNA hybridization.

According to still further features in the described preferred embodiments the at least one oligonucleotide is designed and configured for RNA  
20 hybridization.

According to yet a further aspect of the present invention there is provided a method of annotating biomolecular sequences, the method comprising: (a) computationally clustering the biomolecular sequences according to a progressive homology range, to thereby generate a plurality of  
25 clusters each being of a predetermined homology of the homology range; and (b) assigning at least one ontology to each cluster of the plurality of clusters, the at least one ontology being: (i) derived from an annotation preassociated with at least one biomolecular sequence of each cluster; and/or (ii) generated from analysis of the at least one biomolecular sequence of each cluster thereby  
30 annotating biomolecular sequences.



According to still a further aspect of the present invention there is provided a system for generating a database of annotated biomolecular sequences, the system comprising a processing unit, the processing unit executing a software application configured for: (a) clustering the biomolecular sequences according to a progressive homology range, to thereby generate a plurality of clusters each being of a predetermined homology of the homology range; and (b) assigning at least one ontology to each cluster of the plurality of clusters, the at least one ontology being: (i) derived from an annotation preassociated with at least one biomolecular sequence of each cluster; and/or (ii) generated from analysis of the at least one biomolecular sequence of each cluster, to thereby annotate the biomolecular sequences; and (c) storing sequence annotations and sequence information of the annotated biomolecular sequences, thereby generating the database of annotated biomolecular sequences.

According to still a further aspect of the present invention there is provided a computer readable storage medium comprising a database stored in a retrievable manner, the database including sequence information as set forth in files:

“Transcripts\_nucleotide\_seqs\_part1”  
“Transcripts\_nucleotide\_seqs\_part2”  
“Transcripts\_nucleotide\_seqs\_part3.new”  
“Transcripts\_nucleotide\_seqs\_part4” and/or  
“Protein.seqs”

provided in CD-ROMs 1 and/or 2 enclosed herewith, and sequence ontological annotations in #GO\_P, #GO\_F, #GO\_C annotation categories in file “Summary\_table.new” of CD-ROM3 enclosed herewith.

According to still further features in the described preferred embodiments the biomolecular sequences are selected from the group consisting of polynucleotide sequences and polypeptide sequences.

According to still further features in the described preferred embodiments the homology range is between 99 % - 35 %.

According to still further features in the described preferred embodiments the analysis of the at least one biomolecular sequence includes literature text  
5 mining.

According to still further features in the described preferred embodiments the analysis of the at least one biomolecular sequence includes cellular localization prediction.

According to still further features in the described preferred embodiments  
10 the analysis of the at least one biomolecular sequence includes homology analysis.

According to still further features in the described preferred embodiments the at least one ontology is selected from the group consisting of molecular biology, microbiology, developmental biology, immunology, virology,  
15 biochemistry, physiology, pharmacology, medicine, bioinformatics, cell biology, endocrinology, structural biology, mathematics, chemistry, medicine, plant sciences, neurology, genetics, zoology, ecology, genomics, cheminformatics, computer sciences, statistics, physics and artificial intelligence.

According to still further features in the described preferred embodiments  
20 the ontology includes a subontology.

According to still further features in the described preferred embodiments the method further comprising scoring the at least one ontology assigned to a cluster of the plurality of clusters according to: (i) a degree of homology characterizing the cluster; and (ii) relevance of annotation to information  
25 obtained from literature text mining.

According to still further features in the described preferred embodiments the system further comprising scoring the at least one ontology assigned to a cluster of the plurality of clusters according to: (i) a degree of homology characterizing the cluster; and (ii) relevance of annotation to information  
30 obtained from literature text mining.

According to still further features in the described preferred embodiments the method further comprising generating a sequence profile to each cluster of the plurality of clusters following step (b).

According to still further features in the described preferred  
5 embodiments the system further comprising generating a sequence profile to each cluster of the plurality of clusters following step (b).

According to still a further aspect of the present invention there is provided a computer readable storage medium, comprising a database stored in a retrievable manner, the database including biomolecular sequence information  
10 as set forth in files:

“Transcripts\_nucleotide\_seqs\_part1”

“Transcripts\_nucleotide\_seqs\_part2”

“Transcripts\_nucleotide\_seqs\_part3.new”

“Transcripts\_nucleotide\_seqs\_part4” and/or

15 “Protein.seqs”

provided in CD-ROMs 1 and/or 2 enclosed herewith, and biomolecular sequence annotations as set forth in file “Summary\_table.new” of CD-ROM 3 enclosed herewith.

According to still a further aspect of the present invention there is  
20 provided a method of diagnosing colon cancer in a subject, the method comprising identifying in the subject the presence or absence of a biomolecular sequence selected from the group consisting of SEQ ID NOs: 4, 39, 24-28, 35-38, 12 and 29-31 wherein presence of the biomolecular sequence indicates colon cancer in the subject.

25 According to still a further aspect of the present invention there is provided method of diagnosing lung cancer in a subject, the method comprising identifying in the subject the presence or absence of a biomolecular sequence selected from the group consisting of SEQ ID NOs: 15, 18, 21 and 32 wherein  
{ presence of the biomolecular sequence indicates lung cancer in the subject.

30 According to still a further aspect of the present invention there is

provided a method of diagnosing Ewing sarcoma in a subject, the method comprising identifying in the subject the presence or absence of a biomolecular sequence as set forth in SEQ ID NO: 7, wherein presence of the biomolecular sequence indicates Ewing sarcoma in the subject.

5 According to still a further aspect of the present invention there is provided a computer readable storage medium comprising data stored in a retrievable manner, the data including sequence information of differentially expressed biomolecular sequences as set forth in files:

“Transcripts\_nucleotide\_seqs\_part1”  
10 “Transcripts\_nucleotide\_seqs\_part2”  
“Transcripts\_nucleotide\_seqs\_part3.new”  
“Transcripts\_nucleotide\_seqs\_part4” and  
“Protein.seqs”

provided in CD-ROMs 1 and/or 2 enclosed herewith, and sequence  
15 annotations as set forth in annotation categories “SA” and “RA”, in the file “Summary\_table.new” of CD-ROM3 enclosed herewith.

According to still a further aspect of the present invention there is provided a computer readable storage medium comprising data stored in a retrievable manner, the data including sequence information of biomolecular  
20 sequences exhibiting gain of function or loss of function as set forth in files:  
“Transcripts\_nucleotide\_seqs\_part1”

“Transcripts\_nucleotide\_seqs\_part2”  
“Transcripts\_nucleotide\_seqs\_part3.new”  
“Transcripts\_nucleotide\_seqs\_part4” and  
25 “Protein.seqs”

provided in CD-ROMs 1 and/or 2 enclosed herewith, and sequence annotations as set forth in annotation category “DN”, in the file “Summary\_table.new” of CD-ROM3 enclosed herewith.

According to still further features in the described preferred embodiments  
30 the database further includes information pertaining to generation of the data

and potential uses of the data.

According to still further features in the described preferred embodiments the medium is selected from the group consisting of a magnetic storage medium, an optical storage medium and an optico-magnetic storage medium.

5        According to still further features in the described preferred embodiments the database further includes information pertaining to gain and/or loss of function of the differentially expressed mRNA splice variants or polypeptides encoded thereby.

10        The present invention successfully addresses the shortcomings of the presently known configurations by providing methods and systems useful for systematically annotating biomolecular sequences.

15        Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Although methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present invention, suitable methods and materials are described below. In addition, the materials, methods, and examples are illustrative only and not intended to be limiting.

**BRIEF DESCRIPTION OF THE DRAWINGS**

The invention is herein described, by way of example only, with reference to the accompanying drawings. With specific reference now to the drawings in detail, it is stressed that the particulars shown are by way of example and for purposes of illustrative discussion of the preferred  
5 embodiments of the present invention only, and are presented in the cause of providing what is believed to be the most useful and readily understood description of the principles and conceptual aspects of the invention. In this regard, no attempt is made to show structural details of the invention in more  
10 detail than is necessary for a fundamental understanding of the invention, the description taken with the drawings making apparent to those skilled in the art how the several forms of the invention may be embodied in practice.

In the drawings:

FIG. 1a illustrates a system designed and configured for generating a  
15 database of annotated biomolecular sequences according to the teachings of the present invention.

FIG. 1b illustrates a remote configuration of the system described in Figure 1a.

FIG. 2 illustrates a gastrointestinal tissue hierarchy dendrogram generated  
20 according to the teachings of the present invention.

FIG. 3 is a scheme illustrating multiple alignment of alternatively spliced expressed sequences with a genomic sequence including 3 exons (A, B and C) and two introns. Two alternative splicing events are described; One from the donor site, which involves an AB junction, between donor and proximal  
25 acceptor and an AC junction, between donor and distal acceptor; A Second alternative splicing event is described from the acceptor site, which involves AC junction, between distal donor and acceptor and BC junction, between proximal donor and acceptor.

FIG. 4 is a tissue hierarchy dendrogram generated according to the  
30 teachings of the present invention. The higher annotation levels are marked

with a single number, i.e., 1-16. The lower annotation levels are marked within the relevant category as one – four numbers after the point (e.g. 4. genitourinary system; 4.2 genital system; 4.2.1 women genital system; 4.2.1.1 cervix).

5        FIG. 5 is a graph illustrating a correlation between LOD scores of textual information analysis and accuracy of ontological annotation prediction. Results are based on self-validation studies. Only predictions made with LOD scores above 2 were evaluated and used for GO annotation process.

10        FIGs. 6a-c are histograms showing the distribution of proteins (closed squares) and contigs (opened squares) from Ensembl version 1.0.0 in the major nodes of three GO categories - cellular component (Figure 6a), molecular function (Figure 6b), and biological process (figure 6c).

15        FIG. 7 illustrates results from RT-PCR analysis of the expression pattern of the AA535072 (SEQ ID NO: 39) colorectal cancer-specific transcript. The following cell and tissue samples were tested: B - colon carcinoma cell line SW480 (ATCC-228); C - colon carcinoma cell line SW620 (ATCC-227); D - colon carcinoma cell line colo-205 (ATCC-222). Colon normal tissue indicates a pool of 10 different samples, (Biochain, cat no A406029). The adenocarcinoma sample represents a pool of spleen, lung, stomach and kidney  
20        adenocarcinomas, obtained from patients. Each of the tissues (i.e., colon carcinoma samples Duke's A-D; and normal muscle, pancreas, breast, liver, testis, lung, heart, ovary, thymus, spleen kidney, placenta, stomach, brain) were obtained from 3-6 patients and pooled.

25        FIG. 8 illustrates results from RT-PCR analysis of the expression pattern of the AA513157 (SEQ ID NO: 7) Ewing sarcoma specific transcript. The (+) or (-) symbols, indicate presence or absence of reverse transcriptase in the reaction mixture. A molecular weight standard is indicated by M. Tissue samples (i.e., Ewing sarcoma samples, spleen adenocarcinoma, brain, prostate and thymus) were obtained from patients. The Ln-CAP human prostatic  
30        adenocarcinoma cell line was obtained from the ATCC (Manassas, VA).

FIG. 9 is an autoradiogram of a northern blot analysis depicting tissue distribution and expression levels of AA513157 (SEQ ID NO: 7) Ewing sarcoma specific transcript. Arrows indicate the molecular weight of 28S and 18S ribosomal RNA subunits. The indicated tissue samples were obtained from patients and SK-ES-1 - Ewing sarcoma cell-line was obtained from the ATCC (CRL-1427).

FIG. 10 illustrates results from semi quantitative RT-PCR analysis of the expression pattern of the AA469088 (SEQ ID NO: 40) colorectal specific transcript. Colon normal was obtained from Biochain, cat no: A406029. The adenocarcinoma sample represents a pool of spleen, lung, stomach and kidney adenocarcinomas, obtained from patients. Each of all other tissues (i.e., colon carcinoma samples Duke's A-D; and normal thymus, spleen, kidney, placenta, stomach, brain) were obtained from 3-6 patients and pooled.

FIG. 11 is a histogram depicting Real-Time RT-PCR quantification of copy number, of a lung specific transcript, (SEQ ID NO: 15). Amplification products obtained from the following tissues were quantified; normal salivary gland from total RNA (Clontech, cat no:64110-1); lung normal from pooled adult total RNA (BioChain, cat no:A409363); lung tumor squamos cell carcinoma (Clontech, cat no:64013-1); lung tumor squamos cell carcinoma (BioChain, cat no:A409017); pooled lung tumor squamos cell carcinoma (BioChain, cat no: A411075); moderately differentiated squamos cell carcinoma (BioChain, cat no: A409091); well differentiated squamos cell carcinoma (BioChain, cat no: A408175); pooled adenocarcinoma (BioChain, cat no: A411076); moderately differentiated alveolus cell carcinoma (BioChain, cat no: A409089); non-small cell lung carcinoma cell line H1299; The following normal and tumor samples were obtained from patients: normal lung (internal number-CG-207N), lung carcinoma (internal number-CG-72), squamos cell carcinoma (internal number-CG-196), squamos cell carcinoma (internal number-CG-207), lung adenocarcinoma (internal number-CG-120), lung adenocarcinoma (internal number-CG-160). Copy number was



normalized to the levels of expression of the housekeeping genes Proteasome 26S subunit (dark columns) and GADPH (bright columns).

FIG. 12 is a histogram depicting Real-Time RT-PCR quantification of copy number, of the lung specific transcript (SEQ ID NO: 32). Amplification products obtained from the following tissues and cell-lines were quantified; lung normal from pooled adult total RNA (BioChain, cat no:A409363); lung tumor squamos cell carcinoma (Clontech, cat no:64013-1); lung tumor squamos cell carcinoma (BioChain, cat no:A409017); pooled lung tumor squamos cell carcinoma (BioChain, cat no: A411075); moderately differentiated squamos cell carcinoma (BioChain, cat no: A409091); well differentiated squamos cell carcinoma (BioChain, cat no: A408175); pooled adenocarcinoma (BioChain, cat no: A411076); moderately differentiated alveolus cell carcinoma (BioChain, cat no: A409089); non-small cell lung carcinoma cell line H1299; The following normal and tumor samples were obtained from patients: normal lung (internal number-CG-207N), lung carcinoma (internal number-CG-72), squamos cell carcinoma (internal number-CG-196), squamos cell carcinoma (internal number-CG-207), lung adenocarcinoma (internal number-CG-120), lung adenocarcinoma (internal number-CG-160).. Copy number was normalized to the levels of expression of the housekeeping genes Proteasome 26S subunit (dark columns) and GADPH (bright columns).

FIG. 13 is a histogram depicting Real-Time RT-PCR quantification of copy number, of the lung specific transcript (SEQ ID NO: 18). Amplification products obtained from the following tissues and cell-lines were quantified; lung normal from pooled adult total RNA (BioChain, cat no:A409363); lung tumor squamos cell carcinoma (Clontech, cat no:64013-1); lung tumor squamos cell carcinoma (BioChain, cat no:A409017); pooled lung tumor squamos cell carcinoma (BioChain, cat no: A411075); moderately differentiated squamos cell carcinoma (BioChain, cat no: A409091); well differentiated squamos cell carcinoma (BioChain, cat no: A408175); pooled

adenocarcinoma (BioChain, cat no: A411076); moderately differentiated alveolus cell carcinoma (BioChain, cat no: A409089); non-small cell lung carcinoma cell line H1299; The following normal and tumor samples were obtained from patients: normal lung (internal number-CG-207N), lung carcinoma (internal number-CG-72), squamos cell carcinoma (internal number-CG-196), squamos cell carcinoma (internal number-CG-207), lung adenocarcinoma (internal number-CG-120), lung adenocarcinoma (internal number-CG-160). Copy number was normalized to the levels of expression of the housekeeping genes Proteasome 26S subunit (dark columns) and GADPH (bright columns).

FIG. 14 is a histogram depicting Real-Time RT-PCR quantification of copy number, of a lung specific transcript (SEQ ID NO: 21). Amplification products obtained from the following tissues and cell-lines were quantified; Samples 1-6 are commercial normal lung samples (BioChain, CDP-061010; A503205, A503384, A503385, A503204, A503206, A409363). Sample 7 is lung well differentiated adenocarcinoma (BioChain, CDP-064004A; A504117). Sample 8 is lung moderately differentiated adenocarcinoma (BioChain, CDP-064004A; A504119). Sample 9 is lung moderately to poorly differentiated adenocarcinoma (BioChain, CDP-064004A; A504116). Sample 10 is lung well differentiated adenocarcinoma (BioChain, CDP-064004A; A504118). Samples 11-16 are lung adenocarcinoma samples obtained from patients. Sample 17 is lung moderately differentiated squamous cell carcinoma (BioChain, CDP-064004B; A503187). Sample 18 is lung squamous cell carcinoma (BioChain, CDP-064004B; A503386). Samples 20-21 are lung moderately differentiated squamous cell carcinoma (BioChain, CDP-064004B; A503387, A503383). Sample 22 is lung squamous cell carcinoma pooled (BioChain, CDP-064004B; A411075). Samples 23-26 and sample 31 are lung squamous cell carcinoma obtained from patients. Sample 27 is lung squamous cell carcinoma (Clontech, 64013-1). Sample 28 is lung squamous cell carcinoma (BioChain, A409017). Sample 29 is lung moderately differentiated

squamous cell carcinoma (BioChain, CDP-064004B; A409091). Sample 30 is lung well differentiated squamous cell carcinoma (BioChain, CDP-064004B; A408175). Samples 32-35 are lung small cell carcinoma (BioChain, CDP-064004D; A504115, A501390, A501389, A501391). Sample 36-37 are lung large cell carcinoma (BioChain, CDP-064004C; A504113, A504114). Sample 38 is lung moderately differentiated alveolus cell carcinoma (BioChain, A409089). Sample 39 is lung carcinoma obtained from patient. Sample 40 is lung H1299 non-small cell carcinoma cell line. Sample 41 is normal salivary gland sample (Clontech, 64110-1). Copy number was normalized to the levels of expression of the housekeeping genes Proteasome 26S subunit (dark columns) and GADPH (bright columns).

FIGs. 15a-c are schematic illustrations depicting the methodology undertaken for finding exon-skipping events which are conserved between human and mice genomes. 3,583 exon skipping events were found in the human genome using the methodology described in Sorek (2002) Genome Res. 12:1060-1067. Figure 15a – for 980 of these human exons, a mouse EST spanning the intron which represents the exon-skipping variant was found. Human ESTs are designated in purple. Mouse ESTs are denoted by light blue. Figures 15b-c depict two approaches for identifying exon conservation between mice and human. Figure 15b depicts the identification of mouse ESTs which contain the exon as well as the two flanking exons. Figure 15c illustrates a specific embodiment wherein the exon is absent in the mouse ESTs, in this case the human exon sequence is searched against the intron spanned by the skipping mouse EST on the mouse genome. If a significant conservation (i.e., above 80 %) was found and the alignment spanned the full length of the human exon, the exon was considered conserved.

#### DESCRIPTION OF THE PREFERRED EMBODIMENTS

The present invention is of methods and systems, which can be used for annotating biomolecular sequences. Specifically, the present invention can be

used to identify and annotate differentially expressed biomolecular sequences, such as differentially expressed alternatively spliced sequences.

The principles and operation of the present invention may be better understood with reference to the drawings and accompanying descriptions.

5 Before explaining at least one embodiment of the invention in detail, it is to be understood that the invention is not limited in its application to the details of construction and the arrangement of the components set forth in the following description or illustrated in the drawings. The invention is capable of other embodiments or of being practiced or carried out in various ways. Also, it  
10 is to be understood that the phraseology and terminology employed herein is for the purpose of description and should not be regarded as limiting.

### *Terminology*

As used herein, the term "oligonucleotide" refers to a single stranded or double stranded oligomer or polymer of ribonucleic acid (RNA) or  
15 deoxyribonucleic acid (DNA) or mimetics thereof. This term includes oligonucleotides composed of naturally-occurring bases, sugars and covalent internucleoside linkages (e.g., backbone) as well as oligonucleotides having non-naturally-occurring portions which function similarly. Such modified or substituted oligonucleotides are often preferred over native forms because of  
20 desirable properties such as, for example, enhanced cellular uptake, enhanced affinity for nucleic acid target and increased stability in the presence of nucleases.

The phrase "complementary DNA" (cDNA) refers to the double stranded or single stranded DNA molecule, which is synthesized from a  
25 messenger RNA template.

The term "contig" refers to a series of overlapping sequences with sufficient identity to create a longer contiguous sequence. A plurality of contigs may form a cluster. Clusters are generally formed based upon a specified degree of homology and overlap (e.g., a stringency). The different  
30 contigs in a cluster do not typically represent the entire sequence of the gene,

rather the gene may comprise one or more unknown intervening sequences between the defined contigs.

The term "cluster" refers to a nucleic acid sequence cluster or a protein sequence cluster. The former refers to a group of nucleic acid sequences which share a requisite level of homology and or other similar traits according to a given clustering criterion; and the latter refers to a group of protein sequences which share a requisite level of homology and/or other similar traits according to a given clustering criterion.

A process and/or method to group nucleic acid or protein sequences as such is referred to as clustering, which is typically performed by a clustering (i.e., alignment) application program implementing a cluster algorithm.

As used herein the phrase "biomolecular sequences" refers to amino acid sequences (i.e., peptides, polypeptides) and nucleic acid sequences, which include but are not limited to genomic sequences, expressed sequence tags, contigs, complementary DNA (cDNA) sequences, pre-messenger RNA (mRNA) sequences, and mRNA sequences.

With the presentation of the human genome working draft, data analysis rather than data collection presents the biggest challenge to biologists. Efforts to ascribe biological meaning to genomic data, include the development of advanced wet laboratorial techniques as well as computerized algorithms. While the former are limited due to inaccuracy, time consumption, labor intensiveness and costs the latter are still unfeasible due to the poor organization of on hand sequence databases as well as the composite nature of biological data.

As is further described hereinbelow, the present inventors have developed a computer-based approach for the functional, spatial and temporal analysis of biological data. The present methodology generates comprehensive databases which greatly facilitate the use of available genetic information in both research and commercial applications.

As is further described hereinunder, the present invention encompasses

several novel approaches for annotating biomolecular sequences.

“Annotating” refers to the act of discovering and/or assigning an annotation (i.e., critical or explanatory notes or comment) to a biomolecular sequence of the present invention.

5       The term "annotation" refers to a functional or structural description of a sequence, which may include identifying attributes such as locus name, keywords, Medline references, cloning data, information of coding region, regulatory regions, catalytic regions, name of encoded protein, subcellular localization of the encoded protein, protein hydrophobicity, protein function,  
10       mechanism of protein function, information on metabolic pathways, regulatory pathways, protein-protein interactions and tissue expression profile.

*The ontological annotation approach*

      An ontology refers to the body of knowledge in a specific knowledge domain or discipline such as molecular biology, microbiology, immunology,  
15       virology, plant sciences, pharmaceutical chemistry, medicine, neurology, endocrinology, genetics, ecology, genomics, proteomics, cheminformatics, pharmacogenomics, bioinformatics, computer sciences, statistics, mathematics, chemistry, physics and artificial intelligence.

      An ontology includes domain-specific concepts – referred to herein as  
20       sub-ontologies. A sub-ontology may be classified into smaller and narrower categories.

      The ontological annotation approach of the present invention is effected as follows.

      First, biomolecular sequences are computationally clustered according to  
25       a progressive homology range, thereby generating a plurality of clusters each being of a predetermined homology of the homology range.

      Progressive homology according to this aspect of the present invention is used to identify meaningful homologies among biomolecular sequences and thereby assign new ontological annotations to sequences, which share requisite  
30       levels of homologies. Essentially, a biomolecular sequence is assigned to a

specific cluster if displays a predetermined homology to at least one member of the cluster (i.e., single linkage). As used herein "progressive homology range" refers to a range of homology thresholds, which progress via predetermined increments from a low homology level (e.g. 35 %) to a high homology level (e.g. 99 %). Further description of a progressive homology range is provided in the Examples section which follows.

Following generation of clusters, one or more ontologies are assigned to each cluster. Ontologies are derived from an annotation preassociated with at least one biomolecular sequence of each cluster; and/or generated by analyzing (e.g., text-mining) at least one biomolecular sequence of each cluster thereby annotating biomolecular sequences.

Any annotational information identified and/or generated according to the teachings of the present invention can be stored in a database which can be generated by a suitable computing platform.

Thus, the method according to this aspect of the present invention provides a novel approach for annotating biomolecular sequences even on a scale of a genome, a transcriptome (i.e., the repertoire of all messenger RNA molecules transcribed from a genome) or a proteome (i.e., the repertoire of all proteins translated from messenger RNA molecules). This enables transcriptome-wise comparative analyses (e.g., analyzing chromosomal distribution of human genes) and cross-transcriptome comparative studies (e.g., comparing expressed data across species) both of which may involve various subontologies such as molecular function, biological process and cellular localization.

Biomolecular sequences which can be used as working material for the annotating process according to this aspect of the present invention can be obtained from a biomolecular sequence database. Such a database can include protein sequences and/or nucleic acid sequences derived from libraries of expressed messenger RNA [i.e., expressed sequence tags (EST)], cDNA clones,

contigs, pre-mRNA, which are prepared from specific tissues or cell-lines or from whole organisms.

This database can be a pre-existing publicly available database [i.e., GenBank database maintained by the National Center for Biotechnology Information (NCBI), part of the National Library of Medicine, and the TIGR database maintained by The Institute for Genomic Research, Blocks database maintained by the Fred Hutchinson Cancer Research Center, Swiss-Prot site maintained by the University of Geneva and GenPept maintained by NCBI and including public protein-sequence database which contains all the protein databases from GenBank,] or private databases (i.e., the LifeSeq.<sup>TM</sup> and PathoSeq.<sup>TM</sup> databases available from Incyte Pharmaceuticals, Inc. of Palo Alto, CA). Optionally, biomolecular sequences of the present invention can be assembled from a number of pre-existing databases as described in Example 5 of the Examples section.

Alternatively, the database can be generated from sequence libraries including, but not limited to, cDNA libraries, EST libraries, mRNA libraries and the like.

Construction and sequencing of a cDNA library is one approach for generating a database of expressed mRNA sequences. cDNA library construction is typically effected by tissue or cell sample preparation, RNA isolation, cDNA sequence construction and sequencing.

It will be appreciated that such cDNA libraries can be constructed from RNA isolated from whole organisms, tissues, tissue sections, or cell populations. Libraries can also be constructed from a tissue reflecting a particular pathological or physiological state.

Once raw sequence data is obtained, biomolecular sequences are computationally clustered according to a progressive homology range using one or more clustering algorithms. To obtain progressive clusters, the biomolecular sequences are clustered through single linkage. Namely, a biomolecular sequence belongs to a cluster if this sequence shares a sequence homology



above a certain threshold to one member of the cluster. The threshold increments from a high homology level to a low homology level with a predetermined resolution. Preferably the homology range is selected from 99 % – 35 %.

5 Computational clustering can be effected using any commercially available alignment software including the local homology algorithm of Smith & Waterman, *Adv. Appl. Math.* 2:482 (1981), using the homology alignment algorithm of Needleman & Wunsch, *J. Mol. Biol.* 48:443 (1970), using the search for similarity method of Pearson & Lipman, *Proc. Nat'l. Acad. Sci. USA* 10 85:2444 (1988), or using computerized implementations of algorithms GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package Release 7.0, Genetics Computer Group, 575 Science Dr., Madison, Wis.

Another example of an algorithm which is suitable for sequence alignment is the BLAST algorithm, which is described in Altschul et al., *J. Mol.* 15 *Biol.* 215:403-410 (1990). Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov/>).

Since the present invention requires processing of large amounts of data, sequence alignment is preferably effected using assembly software.

20 A number of commonly used computer software fragment read assemblers capable of forming clusters of expressed sequences, and aligning members of the cluster (individually or as an assembled contig) with other sequences (e.g., genomic database) are now available. These packages include but are not limited to, The TIGR Assembler [Sutton G. et al. (1995) *Genome Science and Technology* 1:9-19], GAP [Bonfield JK. et al. (1995) *Nucleic Acids Res.* 23:4992-4999], CAP2 [Huang X. et al. (1996) *Genomics* 33:21-31], the Genome Construction Manager [Laurence CB. Et al. (1994) *Genomics* 25 23:192-201], Bio Image Sequence Assembly Manager, SeqMan [Swindell SR. and Plasterer JN. (1997) *Methods Mol. Biol.* 70:75-89], and LEADS and 30 GenCarta (Compugen Ltd. Israel).

It will be appreciated that since applying sequence homology analysis on large number of sequences is computationally intensive, local alignment (i.e., the alignment of portions of protein sequences) is preferably effected prior to global alignment (alignment of protein sequences along their entire length), as described in Example 6 of the Examples section.

Once progressive clusters are formed, one or more ontological annotations (i.e., assigning an ontology) are assigned to each cluster.

Systematic and standardized ontological nomenclature is preferably used. Such nomenclature (i.e., keywords) can be obtained from several sources. For example, ontological annotations derived from three main ontologies: molecular function, biological process and cellular component are available from the Gene Ontology Consortium ([www.geneontology.org](http://www.geneontology.org)).

Alternatively a list of homogenized ontological nomenclature can be obtained from AcroMed - a computer generated database of biomedical acronyms and the associated long forms extracted from the recent Medline abstracts (<http://www.expasy.org/tools/>).

Optionally, various conversion tables which link Enzyme Commission number, InterPro protein motifs and SwissProt keywords to gene ontology nodes are also available from [www.geneontology.org](http://www.geneontology.org) and can be used with the present method.

Ontologies, sub ontologies, and their ontological relations (i.e., inherent relation - the sub-ontology "IS THE" ontology or composite relation - the ontology "HAS" the sub ontology) can be organized into various computer data structures such as a tree, a map, a graph, a stack or a list. These may also be presented in various data format such as, text, table, html, or extensible markup language (XML)

Ontologies and/or subontologies assigned to a specific biomolecular sequence can be derived from an annotation, which is preassociated with at least one biomolecular sequence in a cluster generated as described hereinabove.

For example, biomolecular sequences obtained from an annotated database are typically preassociated with an annotation. An "annotated database" refers to a database biomolecular sequences, which are at least partially characterized with respect to functional or structural aspects of the sequence. Examples of annotated databases include but are not limited to:

5 GenBank ([www.ncbi.nlm.nih.gov/GenBank/](http://www.ncbi.nlm.nih.gov/GenBank/)), Swiss-Prot ([www.expasy.ch/sprot/sprot-top.html](http://www.expasy.ch/sprot/sprot-top.html)), GDB ([www.gdb.org/](http://www.gdb.org/)), PIR ([www.mips.biochem.mpg.de/proj/prostseqdb/](http://www.mips.biochem.mpg.de/proj/prostseqdb/)), YDB ([www.mips.biochem.mpg.de/proj/yeast/](http://www.mips.biochem.mpg.de/proj/yeast/)), MIPS ([www.mips.biochem.mpg.de/proj/human](http://www.mips.biochem.mpg.de/proj/human)), HGI ([www.tigr.org/tdb/hgi/](http://www.tigr.org/tdb/hgi/)), Celera Assembled Human Genome ([www.celera.com/products/human\\_ann.cfm](http://www.celera.com/products/human_ann.cfm) and LifeSeq Gold (<http://lifeseqgold.incyte.com>)). Additional specialized annotated databases include annotative information on metabolic

10 (<http://www.genome.ad.jp/kegg/metabolism.html>) and regulatory pathways (<http://www.genome.ad.jp/kegg/regulation.html>), and protein-protein interactions (<http://dip.doe-mbi.ucla.edu/>), etc.

Alternatively, ontologies can be generated from an analysis of at least one biomolecular sequence in each of the clusters of the present invention.

20 Preferably, analysis of the biomolecular sequence is effected by literature text mining. Since manual review of related-literature may be a daunting task, computational extraction of text information is preferably effected.

Thus, the method of the present invention can also process literature and other textual information and utilize processed textual data for generating

25 additional ontological annotations. For example, text information contained in the sequence-related publications and definition lines in sequence records of sequence databases can be extracted and processed. Ontological annotations derived from processed text data are then assigned to the sequences in the corresponding clusters.

Ontological annotations can also be extracted from sequence associated Medical subject heading (MeSH) terms which are assigned to published papers.

Additional information on text mining is provided in Example 7 of the Examples section and is disclosed in "Mining Text Using Keyword Distributions," Ronen Feldman, Ido Dagan, and Haym Hirsh, Proceedings of the 1995 Workshop on Knowledge Discovery in Databases, "Finding Associations in Collections of Text," Ronen Feldman and Haym Hirsh, Machine Learning and Data Mining: Methods and Applications, edited by R. S. Michalski, I. Bratko, and M. Kubat, John Wiley & Sons, Ltd., 1997  
10 "Technology Text Mining, Turning Information Into Knowledge: A White Paper from IBM," edited by Daniel Tkach, Feb. 17, 1998, each of which is fully incorporated herein by reference.

It will be appreciated that text mining may be performed, in this and other embodiments of the present invention, for the text terms extracted from  
15 the definitions of gene or protein sequence records, retrievable from databases such as GenBank and Swiss-Prot and title line, abstract of scientific papers, retrievable from Medline database (e.g., <http://www.ncbi.nlm.nih.gov/PubMed/>).

Computer-dedicated software for biological text analysis is available  
20 from <http://www.expasy.org/tools/>. Examples include, but are not limited to, MedMiner – A software system which extracts and organizes relevant sentences in the literature based on a gene, gene-gene or gene-drug query; Protein Annotator's Assistant – A software system which assists protein annotators in the task of assigning functions to newly sequenced proteins; and  
25 XplorMed – A software system which explores a set of abstracts derived from a bibliographic search in MEDLINE.

Alternatively, assignment of ontological annotations may be effected by analyzing molecular, cellular and/or functional traits of the biomolecular sequences. Prediction of cellular localization may be done using any computer  
30 dedicated software. For example prediction of cellular localization can be done

using the ProLoc (Einat Hazkani-Covo, Erez Levanon, Galit Rotman, Dan Graur and Amit Novik, a manuscript submitted for publication) computational platform. This software is capable of predicting the cellular localization of polypeptide sequences based on inherent features, including specific localization signatures, protein domains, amino acid composition, pI and protein length. Other examples for cellular localization prediction softwares include PSORT - Prediction of protein sorting signals and localization sites and TargetP - Prediction of subcellular location, both available from <http://www.expasy.org/tools/>.

Prediction of functional annotations may be effected by motif analysis of the biomolecular sequences of the present invention. Thus for example, by implementing any motif analysis software, which is based on protein homology (see for example, <http://motif.genome.ad.jp/> and <http://www.accelrys.com/products/grailpro/index.html>) it is possible to predict functional motifs of DNA sequences including repeats, promoter sequences and CpG islands and of encoded proteins such as zinc finger and leucine zipper.

Due to the progressive nature of the clusters of the present invention, ontology assignment starts at the highest level of homology. Any biomolecular sequence in the cluster, which shares identical level of homology compared to an ontologically annotated protein in the cluster is assigned the same ontological annotation. This procedure progresses from the highest level of homology to a lower threshold level with a predetermined increment resolution. Newly discovered homologies enable assignment of existing ontological annotations to biomolecular sequences sharing homologous sequences and being previously unannotated or partially annotated (see Examples 5-9 of the Examples section).

Once assignment of an annotation is effected, annotated clusters are disassembled resulting in annotation of each biomolecular sequence of the cluster.

Such annotated biomolecular sequences are then tested for false

annotation. This is effected using the following scoring parameters:

(i) A degree of homology characterizing the progressive cluster – accuracy of the annotation directly correlates with the homology level used for the annotation process (see Examples 7-9 of the Examples section).

5 (ii) Relevance of annotation to information obtained from literature text mining – each assigned ontological annotation which results from literature text mining or functional or cellular prediction is assessed using scoring parameters such as LOD score (For further details see Example 7 of the Examples section).

10 The present invention also enables the use of the homologies identified according to the teachings of the present invention to annotate more sensitively and rapidly a query sequence. Essentially this involves building a sequence profile for each annotated cluster. A profile enables scoring of a biomolecular sequence according to functional domains along a sequence and generally  
15 makes searches more sensitive. Essentially, clustered sequences are also tested for relevance to the cluster based upon shared functional domains and other characteristic sequence features.

Ontologically annotated biomolecular sequences are stored in a database for further use. Additional information on generation and contents of such  
20 databases is provided hereinunder.

Such a database can be used to query functional domains and sequences comprising thereof. Alternatively, the database can be used to query a sequence, and retrieve the compatible annotations.

Although the present methodology can be effected using prior art  
25 systems modified for such purposes, due to the large amounts of data processed and the vast amounts of processing needed, the present methodology is preferably effected using a dedicated computational system.

Thus, according to another aspect of the present invention and as illustrated in Figures 1a-b, there is provided a system for generating a database  
30 of annotated biomolecular sequences.

System 10 includes a processing unit 12, which executes a software application designed and configured for annotating biomolecular sequences, as described hereinabove. System 10 further serves for storing biomolecular sequence information and annotations in a retrievable/searchable database 18.

5 Database 18 further includes information pertaining to database generation.

System 10 may also include a user interface 14 (e.g., a keyboard and/or a mouse, monitor) for inputting database or database related information, and for providing database information to a user.

System 10 of the present invention may be any computing platform  
10 known in the art including but not limited to a personal computer, a work station, a mainframe and the like.

Preferably, database 18 is stored on a computer readable media such as a magnetic optico-magnetic or optical disk.

System 10 of the present invention may be used by a user to query the  
15 stored database of annotations and sequence information to retrieve biomolecular sequences stored therein according to inputted annotations or to retrieve annotations according to a biomolecular sequence query.

It will be appreciated that the connection between user interface 14 and processing unit 12 is bi-directional. Likewise, processing unit 12 and database  
20 18 also share a two-way communication channel, wherein processing unit 12 may also take input from database 18 in performing annotations and iterative annotations. Further, user interface 14 is linked directly to database 18, such a user may dispatch queries to database 18 and retrieve information stored therein. As such, user interface 14 allows a user to compile queries, send  
25 instructions, view querying results and performing specific analyses on the results as needed.

In performing ontological annotations, processing unit 12 may take input from one or more application modules 16. Application module 16 performs a specific operation and produced a relevant annotative input for processing unit  
30 12. For example, application module 16 may perform cellular localization

analysis on a biomolecular sequence query, thereby determining the cellular localization of the encoded protein. Such a functional annotation is then input to and used by processing unit 12. Examples for application software for cellular localization prediction are provided hereinabove.

5        System 10 of the present invention may also be connected to one or more external databases 20. External database 20 is linked to processing unit 12 in a bi-directional manner, similar to the connection between database 18 and processing unit 12. External database 20 may include any background information and/or sequence information that pertains to the biomolecular  
10    sequence query. External database 20 may be a proprietary database or a publicly available database which is accessible through a public network such as the Internet. External database 20 may feed relevant information to processing unit 12 as it effects iterative ontological annotation. External database 20 may also receive and store ontological annotations generated by  
15    processing unit 12. In this case external database 20 may interact with other components of system 10 like database 18.

It will be appreciated that the databases and application modules of system 10 can be directly connected with processing unit 12 and/or user interface 14 as is illustrated in Figure 1a, or such a connection can be achieved  
20    via a network 22, as is illustrated in Figure 1b.

Network 22 may be a private network (e.g., a local area network), a secured network, or a public network (such as the Internet), or a combination of public and private and/or secured networks.

Thus, the present invention provides a well characterized approach for  
25    the systemic annotation of biomolecular sequences. The use of text information analysis, annotation scoring system and robust sequence clustering procedure enables for the first time the creation of the best possible annotations and assignment thereof to a vast number of biomolecular sequences sharing homologous sequences. The availability of ontological annotations for a  
30    significant number of biomolecular sequences from different species can



provide a comprehensive account of sequence, structural and functional information pertaining to the biomolecular sequences of interest.

*The hierarchical annotation approach*

“Hierarchical annotation” refers to any ontology and subontology, which  
5 can be hierarchically ordered. Examples include but are not limited to a tissue expression hierarchy, a developmental expression hierarchy, a pathological expression hierarchy, a cellular expression hierarchy, an intracellular expression hierarchy, a taxonomical hierarchy, a functional hierarchy and so forth.

According to another aspect of the present invention there is provided a  
10 method of annotating biomolecular sequences according to a hierarchy of interest. The method is effected as follows.

First, a dendrogram representing the hierarchy of interest is computationally constructed. As used herein a “dendrogram” refers to a branching diagram containing multiple nodes and representing a hierarchy of  
15 categories based on degree of similarity or number of shared characteristics.

Each of the multiple nodes of the dendrogram is annotated by at least one keyword describing the node, and enabling literature and database text mining, as is further described hereinunder. A list of keywords can be obtained from the GO Consortium ([www.geneontology.org](http://www.geneontology.org)); measures are taken to include as  
20 many keywords, and to include keywords which might be out of date. For example, for tissue annotation (see Figure 4), a hierarchy was built using all available tissue/libraries sources available in the GenBank, while considering the following parameters: ignoring GenBank synonyms, building anatomical hierarchies, enabling flexible distinction between tissue types (normal versus  
25 pathology) and tissue classification levels (organs, systems, cell types, etc.).

It will be appreciated that the dendrogram of the present invention can be illustrated as a graph, a list, a map or a matrix or any other graphic or textual organization, which can describe a dendrogram. An example of a dendrogram illustrating the gastrointestinal tissue hierarchy is provided in Figure 2.

In a second step, each of the biomolecular sequences is assigned to at least one specific node of the dendrogram.

The biomolecular sequences according to this aspect of the present invention can be annotated biomolecular sequences, unannotated biomolecular sequences or partially annotated biomolecular sequences.

Annotated biomolecular sequences can be retrieved from pre-existing annotated databases as described hereinabove.

For example, in GenBank, relevant annotational information is provided in the definition and keyword fields. In this case, classification of the annotated biomolecular sequences to the dendrogram nodes is directly effected. A search for suitable annotated biomolecular sequences is performed using a set of keywords which are designed to classify the biomolecular sequences to the hierarchy (i.e., same keywords that populate the dendrogram)

In cases where the biomolecular sequences are unannotated or partially annotated, extraction of additional annotational information is effected prior to classification to dendrogram nodes. This can be effected by sequence alignment, as described hereinabove. Alternatively, annotational information can be predicted from structural studies. Where needed, nucleic acid sequences can be transformed to amino acid sequences to thereby enable more accurate annotational prediction.

Finally, each of the assigned biomolecular sequences is recursively classified to nodes hierarchically higher than the specific nodes, such that the root node of the dendrogram encompasses the full biomolecular sequence set, which can be classified according to a certain hierarchy, while the offspring of any node represent a partitioning of the parent set.

For example, a biomolecular sequence found to be specifically expressed in "rhabdomyosarcoma", will be classified also to a higher hierarchy level, which is "sarcoma", and then to "Mesenchimal cell tumors" and finally to a highest hierarchy level "Tumor". In another example, a sequence found to be differentially expressed in endometrium cells, will be classified also to a higher

hierarchy level, which is "uterus", and then to "women genital system" and to "genital system" and finally to a highest hierarchy level "genitourinary system". The retrieval can be performed according to each one of the requested levels.

Since annotation of publicly available databases is at times unreliable, newly annotated biomolecular sequences are confirmed using computational or laboratory approaches as is further described hereinbelow.

It will be appreciated that once temporal or spatial annotations of sequences are established using the teachings of the present invention, it is possible to identify those sequences, which are differentially expressed (i.e., exhibit spatial or temporal pattern of expression in diverse cells or tissues). Such sequences are assigned to only a portion of the nodes, which constitute the hierarchical dendrogram.

Changes in gene expression are important determinants of normal cellular physiology, including cell cycle regulation, differentiation and development, and they directly contribute to abnormal cellular physiology, including developmental anomalies, aberrant programs of differentiation and cancer. Accordingly, the identification, cloning and characterization of differentially expressed genes can provide relevant and important insights into the molecular determinants of processes such as growth, development, aging, differentiation and cancer. Additionally, identification of such genes can be useful in development of new drugs and diagnostic methods for treating or preventing the occurrence of such diseases.

Newly annotated sequences identified according to the present invention are tested under physiological conditions (i.e., temperature, pH, ionic strength, viscosity, and like biochemical parameters which are compatible with a viable organism, and/or which typically exist intracellularly in a viable cultured yeast cell or mammalian cell). This can be effected using various laboratory approaches such as, for example, FISH analysis, PCR, RT-PCR, southern blotting, northern blotting, electrophoresis and the like (see Examples 13-20 of

the Examples section) or more elaborate approaches which are detailed in the Background section.

It will be appreciated that true involvement of differentially expressed genes in a biological process is better confirmed using an appropriate cell or animal model, as further described hereinunder.

Although the present methodology can be effected using prior art systems modified for such purposes, due to the large amounts of data processed and the vast amounts of processing needed, the present methodology is preferably effected using a dedicated computational system.

Such a system is described hereinabove. The system includes a processing unit which executes a software application designed and configured for hierarchically annotating biomolecular sequences as described hereinabove. The system further serves for storing biomolecular sequence information and annotations in a retrievable/searchable database.

The hierarchical annotation approach enables to assign an appropriate annotation level even in cases where expression is not restricted to a specific tissue type or cell type. For example, different expressed sequences of a single contig which are annotated as being expressed in several different tissue types of a single specific organ or a specific system, are also annotated by the present invention to a higher hierarchy level thus denoting association with the specific organ or system. In such cases using keywords alone would not efficiently identify differentially expressed sequences. Thus for example, a sequence found to be expressed in sarcoma, Ewing sarcoma tumors, pnet, rhabdomyosarcoma, liposarcoma and mesenchymal cell tumors, can not be assigned to specific sarcomas, but still can be annotated as mesenchymal cell tumor specific. Using this hierarchical annotation approach in combination with advanced sequence clustering and assembly algorithms, capable of predicting alternative splicing, may facilitate a simple and rapid identification of gene expression patterns.

*Annotation of differentially expressed alternatively spliced sequences*

Although numerous methods have been developed to identify differentially expressed genes, none of these addressed splice variants, which occur in over 50 % of human genes. Given the common sequence features of splice variants it is very difficult to identify splice variants which expression is differential, using prior art methodologies. Therefore assigning unique sequence features to differentially expressed splice variants may have an important impact to the understanding of disease development and may serve as valuable markers to various pathologies.

Thus, according to another aspect of the present invention there is provided a method of identifying sequence features unique to differentially expressed mRNA splice variants. The method is effected as follows.

First, unique sequence features are computationally identified in identified splice variants of alternatively spliced expressed sequences.

As used herein the phrase "splice variants" refers to naturally occurring nucleic acid sequences and proteins encoded therefrom which are products of alternative splicing. Alternative splicing refers to intron inclusion, exon exclusion, or any addition or deletion of terminal sequences, which results in sequence dissimilarities between the splice variant sequence and the wild-type sequence.

Although most alternatively spliced variants result from alternative exon usage, some result from the retention of introns not spliced-out in the intermediate stage of RNA transcript processing.

As used herein the phrase "unique sequence features" refers to donor/acceptor concatenations (i.e., exon-exon junctions), intron sequences, alternative exon sequences and alternative polyadenylation sequences.

Once a unique sequence feature is identified, the expression pattern of the splice variant is determined. If the splice variant is differentially expressed then the unique feature thereof is annotated accordingly.

Alternatively spliced expressed sequences of this aspect of the present

invention, can be retrieved from numerous publicly available databases. Examples include but are not limited to ASDB - an alternative splicing database generated using GenBank and Swiss-Prot annotations (<http://cbcg.nerisc.gov/asdb>), AsMamDB - a database of alternative splices in human, mouse and rat (<http://166.111.30.65/ASMAMDB.html>), Alternative splicing database - a database of alternative splices from literature ([http://cgsigm.cshl.org/new\\_alt\\_exon\\_db2/](http://cgsigm.cshl.org/new_alt_exon_db2/)), Yeast intron database - Database of intron in yeast ([http://www.cse.ucsc.edu/research/compbio/yeast\\_introns.html](http://www.cse.ucsc.edu/research/compbio/yeast_introns.html)), The Intronerator - alternative splicing in *C. elegans* based on analysis of EST data (<http://www.cse.ucsc.edu/~kent/intronerator>), ISIS - Intron Sequence Information System including a section of human alternative splices (<http://isis.bit.uq.edu.au/>), TAP - Transcript Assembly Program result of alternative splicing (<http://stl.wustl.edu/~zkan/TAP/>) and HASDB - database of alternative splices detected in human EST data.

Additionally, alternative splicing sequence data utilized by this aspect of the present invention can be obtained by any of the following bioinformatical approaches.

(i) *Genomically aligned ESTs* - the method identifies ESTs which come from the same gene and looks for differences between them that are consistent with alternative splicing, such as large insertion or deletion in one EST. Each candidate splice variant can be further assessed by aligning the ESTs with respective genomic sequence. This reveals candidate exons (i.e., matches to the genomic sequence) separated by candidate splices (i.e., large gaps in the EST-genomic alignment). Since intronic sequences at splice junctions (i.e., donor/acceptor concatenations) are highly conserved (essentially 99.24 % of introns have a GT-AG at their 5' and 3' ends, respectively) sequence data can be used to verify candidate splices [Burset et al. (2000) Nucleic Acids Res. 28:4364-75 LEADS module [Shoshan, et al, Proceeding of SPIE (eds. M.L. Bittner, Y. Chen, A.N. Dorsel, E.D. Dougherty) Vol. 4266, pp.

86-95 (2001).;R. Sorek, G. Ast, D. Graur, Genome Res. In press; Compugen Ltd. US patent application 09/133,987].

(ii) *Identification based on intron information* – The method creates a database of individual intron sequences annotated in GenBank and utilizes such sequences to search for EST sequences which include the intronic sequences [Croft et al. (2000) Nat. Genet. 24:340-1].

(iii) *EST alignment to expressed sequences* - looks for insertions and deletions in ESTs relative to a set of known mRNAs. Such a method enables to uncover alternatively spliced variants with having to align ESTs with genomic sequence [Brett et al. (2000) FEBS Lett. 474-83-86].

It will be appreciated that in order to avoid false positive identification of novel splice isoforms, a set of filters is applied. For example, sequences are filtered to exclude EST having sequence deviations, such as chimerism, random variation in which a given EST sequence or potential vector contamination at the ends of an EST.

Filtering can be effected by aligning ESTs with corresponding genomic sequences. Chimeric ESTs can be easily excluded by requiring that each EST aligns completely to a single genomic locus. Genomic location found by homology search and alignment can often be checked against radiation hybrid mapping data [Muneer et al (2002) Genomic 79:344-8]. Furthermore, since the genomic regions which align with an EST sequence correspond to exon sequences and alignment gaps correspond to introns, the putative splice sites at exon/intron boundaries can be confirmed. Because splice donor and acceptor sites primarily reside within the intron sequence, this methodology can provide validation which is independent of the EST evidence. Reverse transcriptase artifacts or other cDNA synthesis errors may also be filtered out using this approach. Improper inclusion of genomic sequence in ESTs can also be excluded by requiring pairs of mutually exclusive splices in different ESTs.

Additionally, it will be appreciated that observing a given splice variant in one EST but not in a second EST may be insufficient, as the latter can be an

un-spliced EST rather than a biological significant intron inclusion. Therefore measures are taken to focus on mutually exclusive splice variants, two different splice variants observed in different ESTs, which overlap in a genomic sequence. A more stringent filtering may be applied by requiring two splice variants to share one splice site but differ in another. Another filter which can be used to identify true splicing events is sequence conservation. Essentially, exons and the borders of human introns which are identified in mice genome and/or supported by mouse ESTs are considered true splicing events (see Example 21 of the Examples section).

Once splice variants are identified, identification of unique sequence features therewithin can be effected computationally by identifying insertions, deletions and donor-acceptor concatenations in ESTs relative to mRNA and preferably genomic sequences.

As mentioned hereinabove, once alternatively spliced sequences (having unique sequence features) are identified, determination of their expression patterns is effected in order to assign an annotation to the unique sequence feature thereof.

Expression pattern identification may be effected by qualifying annotations which are preassociated with the alternatively spliced expressed sequences, as described hereinabove. This can be accomplished by scoring the annotations. For example scoring pathological expression annotations can be effected according to: (i) prevalence of the alternatively spliced expressed sequences in normal tissues; (ii) prevalence of the alternatively spliced expressed sequences in pathological tissues; (iii) prevalence of the alternatively spliced expressed sequence in total tissues; and (iv) number of tissues and/or tissue types expressing the alternatively spliced expressed sequences.

Alternatively, identifying the expression pattern of the alternatively spliced expressed sequences of the present invention, is accomplished by identifying the unique sequence feature thereof. This can be effected by any hybridization-based technique known in the art, such as northern blot, dot blot,



RNase protection assay, RT-PCR and the like.

To this end oligonucleotides probes, which are substantially homologous to nucleic acid sequences that flank and/or extend across the unique sequence features of the alternatively spliced expressed sequences of the present invention are generated.

Preferably, oligonucleotides which are capable of hybridizing under stringent, moderate or mild conditions, as used in any polynucleotide hybridization assay are utilized. Further description of hybridization conditions is provided hereinunder.

Oligonucleotides generated by the teachings of the present invention may be used in any modification of nucleic acid hybridization based techniques, which are further detailed hereinunder. General features of oligonucleotide synthesis and modifications are also provided hereinunder.

Aside from being useful in identifying specific splice variants, oligonucleotides generated according to the teachings of the present invention may also be widely used as diagnostic, prognostic and therapeutic agents in a variety of disorders which are associated with specific splice variants.

Regulation of splicing is involved in 15 % of genetic diseases [Krawczak et al. (1992) Hum. Genet. 90:41-54] and may contribute for example to cancer mis-splicing of exon 18 in BRCA1, which is caused by a polymorphism in an exonic enhancer [Liu et al. (2001) Nature Genet. 27:55-58].

Thus, oligonucleotides generated according to the teachings of the present invention can be included in diagnostic kits. For example, oligonucleotides sets pertaining to a specific disease associated with differential expression of an alternatively spliced transcript can be packaged in a one or more containers with appropriate buffers and preservatives along with suitable instructions for use and used for diagnosis or for directing therapeutic treatment. Additional information on such diagnostic kits is provided hereinunder.

It will be appreciated that an ability to identify alternatively spliced

sequences, also facilitates identification of the various products of alternative splicing.

Recent studies indicate that most alternative splicing events result in an altered protein product [International human genome sequencing consortium  
5 (2001) Nature 409:860-921; Modrek et al. (2001) Nucleic Acids Res. 29:2850-2859]. The majority of these changes appear to have a functional relevance (i.e., up-regulating or down-regulating activity), such as the replacement of the amino or carboxyl terminus, or in-frame addition and removal of a functional domain. For example, alternative splicing can lead to the use of a different site  
10 for translation initiation (i.e., alternative initiation), a different translation termination site due to a frameshift (i.e., truncation or extension), or the addition or removal of a stop codon in the alternative coding sequence (i.e., alternative termination). Additionally, alternative splicing can change an internal sequence region due to an in-frame insertion or deletion. One example  
15 of the latter is the new FC receptor  $\beta$ -like protein, whose C-terminal transmembrane domain and cytoplasmic tail, which is important for signal transduction in this class of receptors, is replaced with a new transmembrane domain and tail by alternative polyadenylation. Another example is the truncated Growth Hormone Receptor which lacks most of its intracellular  
20 domain and has been shown to heterodimerize with the full-length receptor, thus causing inhibition of signaling by Growth Hormone [Ross, R.J.M., Growth hormone & IGF Research, 9:42-46, (1999)].

Thus, assigning a unique sequence feature to a functionally altered splice variant enables identification of such variants. As used herein the phrase  
25 "functionally altered splice variants" refers to alternatively spliced expressed sequences, which protein products exhibit gain of function or loss of function or modification of the original function.

As used herein the phrase "gain of function" refers to any alternative splicing product, which exhibits increased functionality as compared to the wild  
30 type gene product.

As used herein the phrase "loss of function" refers to any alternative splicing product, which exhibits reduced function as compared to the wild type gene product including any reduction in function, total absence of function or dominant negative function.

5 As used herein the phrase "dominant negative" refers to the dominant effect of a splice variant on the activity of wild type mRNA. For example, a protein product of an altered splice variant may bind a wild type target protein without enzymatically activating it (e.g., receptor dimmers), thus blocking and preventing the active enzymes from binding and activating the target protein.

10 As used herein the phrase "functional domain" refers to a region of a polypeptide, which displays a particular function. This function may give rise to a biological, chemical, or physiological consequence which may be reversible or irreversible and which may include protein-protein interactions (e.g., binding interactions) involving the functional domain, a change in the conformation or a  
15 transformation into a different chemical state of the functional domain or of molecules acted upon by the functional domain, the transduction of an intracellular or intercellular signal, the regulation of gene or protein expression, the regulation of cell growth or death, or the activation or inhibition of an immune response.

20 Identification of putative functionally altered splice variants, according to this aspect of the present invention, can be effected by identifying sequence deviations from functional domains of wild-type gene products.

Identification of functional domains can be effected by comparing a wild-type gene product with a series of profiles prepared by alignment of well  
25 characterized proteins from a number of different species. This generates a consensus profile, which can then be matched with the query sequence. Examples of programs suitable for such identification include, but are not limited to, InterPro Scan - Integrated search in PROSITE, Pfam, PRINTS and other family and domain databases; ScanProsite - Scans a sequence against  
30 PROSITE or a pattern against SWISS-PROT and TrEMBL; MotifScan - Scans

a sequence against protein profile databases (including PROSITE); Frame-ProfileScan - Scans a short DNA sequence against protein profile databases (including PROSITE); Pfam HMM search - scans a sequence against the Pfam protein families database; FingerPRINTScan - Scans a protein sequence against the PRINTS Protein Fingerprint Database; FPAT - Regular expression searches in protein databases; PRATT - Interactively generates conserved patterns from a series of unaligned proteins; PPSEARCH - Scans a sequence against PROSITE (allows a graphical output); at EBI; PROSITE scan - Scans a sequence against PROSITE (allows mismatches); at PBIL; PATTINPROT - Scans a protein sequence or a protein database for one or several pattern(s); at PBIL; SMART - Simple Modular Architecture Research Tool; at EMBL; TEIRESIAS - Generate patterns from a collection of unaligned protein or DNA sequences; at IBM, all available from <http://www.expasy.org/tools/>.

It will be appreciated that functionally altered splice variants may also include a sequence alteration at a post-translation modification consensus site, such as, for example, a tyrosine sulfation site, a glycosylation site, etc. Examples of post-translational modification prediction softwares include but are not limited to: SignalP - Prediction of signal peptide cleavage sites; ChloroP - Prediction of chloroplast transit peptides; MITOPROT - Prediction of mitochondrial targeting sequences; Predotar - Prediction of mitochondrial and plastid targeting sequences; NetOGlyc - Prediction of type O-glycosylation sites in mammalian proteins; DictyOGlyc - Prediction of GlcNAc O-glycosylation sites in Dictyostelium; YinOYang - O-beta-GlcNAc attachment sites in eukaryotic protein sequences; big-PI Predictor - GPI Modification Site Prediction; DGPI - Prediction of GPI-anchor and cleavage sites (Mirror site); NetPhos - Prediction of Serine, Threonine and Tyrosine phosphorylation sites in eukaryotic proteins; NetPicoRNA - Prediction of protease cleavage sites in picornaviral proteins; NMT - Prediction of N-terminal N-myristoylation; Sulfinator - Prediction of tyrosine sulfation sites all available from <http://www.expasy.org/tools/>.

Once putative functionally altered splice variants are identified, they are validated by experimental verification and functional studies, using methodologies well known in the art.

The Examples section which follows illustrates identification and annotation of splice variants. Identified and annotated sequences are contained within the enclosed CD-ROMs 1-3. Some of these sequences represent (i.e., are transcribed from) entirely new splice variants, while others represent new splice variants of known sequences. In any case, the sequences contained in the enclosed CD-ROM are novel in that they include previously undisclosed sequence regions in the context of a known gene or an entirely new sequence in the context of an unknown gene.

The nucleic acids of the invention can be "isolated" or "purified." In the event the nucleic acid is genomic DNA, it is considered "isolated" when it does not include coding sequence(s) of a gene or genes immediately adjacent thereto in the naturally occurring genome of an organism; although some or all of the 5' or 3' non-coding sequence of an adjacent gene can be included. For example, an isolated nucleic acid (DNA or RNA) can include some or all of the 5' or 3' non-coding sequence that flanks the coding sequence (e.g., the DNA sequence that is transcribed into, or the RNA sequence that gives rise to, the promoter or an enhancer in the mRNA). For example, an isolated nucleic acid can contain less than about 5 kb (e.g., less than about 4kb, 3 kb, 2 kb, 1kb, 0.5kb, or 0.1 kb) of the 5' and/or 3' sequence that naturally flanks the nucleic acid molecule in a cell in which the nucleic acid naturally occurs. In the event the nucleic acid is RNA or mRNA, it is "isolated" or "purified" from a natural source (e.g., a tissue) or a cell culture when it is substantially free of the cellular components with which it naturally associates in the cell and, if the cell was cultured, the cellular components and medium in which the cell was cultured (e.g., when the RNA or mRNA is in a form that contains less than about 20%, 10%, 5%, 1%, or less, of other cellular components or culture medium). When chemically synthesized, a nucleic acid (DNA or RNA) is "isolated" or

“purified” when it is substantially free of the chemical precursors or other chemicals used in its synthesis (e.g., when the nucleic acid is in a form that contains less than about 20%, 10%, 5%, 1%, or less, of the chemical precursors or other chemicals).

5 Variants, fragments, and other mutant nucleic acids are also envisaged by the present invention. As noted above, where a given biomolecular sequence represents a new gene (rather than a new splice variant of a known gene), the nucleic acids of the invention include the corresponding genomic DNA and RNA. Accordingly, where a given SEQ ID represents a new gene, variations or  
10 mutations can occur not only in that nucleic acid sequence, but in the coding regions, the non-coding regions, or both, of the genomic DNA or RNA from which it was made.

The nucleic acids of the invention can be double-stranded or single-stranded and can, therefore, either be a sense strand, an antisense strand, or a  
15 portion (i.e., a fragment) of either the sense or the antisense strand. The nucleic acids of the invention can be synthesized using standard nucleotides or nucleotide analogs or derivatives (e.g., inosine, phosphorothioate, or acridine substituted nucleotides), which can alter the nucleic acid’s ability to pair with complementary sequences or to resist nucleases. Indeed, the stability or  
20 solubility of a nucleic acid can be altered (e.g., improved) by modifying the nucleic acid’s base moiety, sugar moiety, or phosphate backbone. For example, the nucleic acids of the invention can be modified as taught by Toulmé [Nature Biotech. 19:17, (2001)] or Faria et al. [Nature Biotech. 19:40-44, (2001)], and the deoxyribose phosphate backbone of nucleic acids can be modified to  
25 generate peptide nucleic acids [PNAs; see Hyrup et al., (1996) Bioorganic & Medicinal Chemistry 4:5-23].

PNAs are nucleic acid “mimics”; the molecule’s natural backbone is replaced by a pseudopeptide backbone and only the four nucleotide bases are retained. This allows specific hybridization to DNA and RNA under conditions  
30 of low ionic strength. PNAs can be synthesized using standard solid phase

peptide synthesis protocols as described, for example by Hyrup et al. (supra) and Perry-O'Keefe et al. [Proc. Natl. Acad. Sci. USA (1996) 93:14670-675]. PNAs of the nucleic acids described herein can be used in therapeutic and diagnostic applications.

5        Moreover, the nucleic acids of the invention include not only protein-encoding nucleic acids per se (e.g., coding sequences produced by the polymerase chain reaction (PCR) or following treatment of DNA with an endonuclease), but also, for example, recombinant DNA that is: (a) incorporated into a vector (e.g., an autonomously replicating plasmid or virus),  
10        (b) incorporated into the genomic DNA of a prokaryote or eukaryote, or (c) part of a hybrid gene that encodes an additional polypeptide sequence (i.e., a sequence that is heterologous to the nucleic acid sequences of the present invention or fragments, other mutants, or variants thereof).

      This aspect of the present invention includes naturally occurring  
15        sequences of the nucleic acid sequences described above, allelic variants (same locus; functional or non-functional), homologs (different locus), and orthologs (different organism) as well as degenerate variants of those sequences and fragments thereof. The degeneracy of the genetic code is well known, and one of ordinary skill in the art will be able to make nucleotide sequences that differ  
20        from the nucleic acid sequences of the present invention but nevertheless encode the same proteins as those encoded by the nucleic acid sequences of the present invention. The variant sequences (e.g., degenerate variants) can be used in the same manner as naturally occurring sequences. For example, the variant DNA sequences of the invention can be incorporated into a vector, into the genomic  
25        DNA of a prokaryote or eukaryote, or made part of a hybrid gene. Moreover, variants (or, where appropriate, the proteins they encode) can be used in the diagnostic assays and therapeutic regimes described below.

      The sequence of nucleic acids of the invention can also be varied to maximize expression in a particular expression system. For example, as few as  
30        one and as many as about 20% of the codons in a given sequence can be altered

to optimize expression in bacterial cells (e.g., *E. coli*), yeast, human, insect, or other cell types (e.g., CHO cells).

The nucleic acids of the invention can also be shorter or longer than those disclosed on CD-ROMs 1 and 2. Where the nucleic acids of the invention encode proteins, the protein-encoding sequences can differ from those represented by specific sequences of file "Protein.seqs" in CD-ROM 2. For example, the encoded proteins can be shorter or longer than those encoded by one of the nucleic acid sequences of the present invention. Nucleotides can be deleted from, or added to, either or both ends of the nucleic acid sequences of the present invention or the novel portions of the sequences that represent new splice variants. Alternatively, the nucleic acids can encode proteins in which one or more amino acid residues have been added to, or deleted from, one or more sequence positions within the nucleic acid sequences.

The nucleic acid fragments can be short (e.g., 15-30 nucleotides). For example, in cases where peptides are to be expressed therefrom such polynucleotides need only contain a sufficient number of nucleotides to encode novel antigenic epitopes. In cases where nucleic acid fragments serve as DNA or RNA probes or PCR primers, fragments are selected of a length sufficient for specific binding to one of the sequences representing a novel gene or a unique portion of a novel splice variant.

Nucleic acids used as probes or primers are often referred to as oligonucleotides, and they can hybridize with a sense or antisense strand of DNA or RNA. Nucleic acids that hybridize to a sense strand (i.e., a nucleic acid sequence that encodes protein, e.g., the coding strand of a double-stranded cDNA molecule) or to an mRNA sequence are referred to as antisense oligonucleotides. Antisense oligonucleotides can be used to specifically inhibit transcription of any of the nucleic acid sequences of the present invention.

Design of antisense molecules must be effected while considering two aspects important to the antisense approach. The first aspect is delivery of the oligonucleotide into the cytoplasm of the appropriate cells, while the second



aspect is design of an oligonucleotide which specifically binds the designated mRNA within cells in a way which inhibits translation thereof.

The prior art teaches of a number of delivery strategies which can be used to efficiently deliver oligonucleotides into a wide variety of cell types [see, for example, Luft (1998) *J Mol Med* 76(2): 75-6; Kronenwett et al. (1998) *Blood* 91(3): 852-62; Rajur et al. (1997) *Bioconjug Chem* 8(6): 935-40; Lavigne et al. (1997) *Biochem Biophys Res Commun* 237(3): 566-71 and Aoki et al. (1997) *Biochem Biophys Res Commun* 231(3): 540-5].

In addition, algorithms for identifying those sequences with the highest predicted binding affinity for their target mRNA based on a thermodynamic cycle that accounts for the energetics of structural alterations in both the target mRNA and the oligonucleotide are also available [see, for example, Walton et al. (1999) *Biotechnol Bioeng* 65(1): 1-9].

Such algorithms have been successfully used to implement an antisense approach in cells. For example, the algorithm developed by Walton et al. enabled scientists to successfully design antisense oligonucleotides for rabbit beta-globin (RBG) and mouse tumor necrosis factor-alpha (TNF alpha) transcripts. The same research group has more recently reported that the antisense activity of rationally selected oligonucleotides against three model target mRNAs (human lactate dehydrogenase A and B and rat gp130) in cell culture as evaluated by a kinetic PCR technique proved effective in almost all cases, including tests against three different targets in two cell types with phosphodiester and phosphorothioate oligonucleotide chemistries.

In addition, several approaches for designing and predicting efficiency of specific oligonucleotides using an in vitro system were also published (Matveeva et al. (1998) *Nature Biotechnology* 16, 1374 - 1375).

Several clinical trials have demonstrated safety, feasibility and activity of antisense oligonucleotides. For example, antisense oligonucleotides suitable for the treatment of cancer have been successfully used (Holmund et al. (1999) *Curr Opin Mol Ther* 1(3):372-85), while treatment of hematological

malignancies via antisense oligonucleotides targeting c-myc gene, p53 and Bcl-2 had entered clinical trials and had been shown to be tolerated by patients [Gerwitz (1999) *Curr Opin Mol Ther* 1(3):297-306].

More recently, antisense-mediated suppression of human heparanase  
5 gene expression has been reported to inhibit pleural dissemination of human cancer cells in a mouse model [Uno et al. (2001) *Cancer Res* 61(21):7855-60].

Thus, the current consensus is that recent developments in the field of antisense technology which, as described above, have led to the generation of highly accurate antisense design algorithms and a wide variety of  
10 oligonucleotide delivery systems, enable an ordinarily skilled artisan to design and implement antisense approaches suitable for downregulating expression of known sequences without having to resort to undue trial and error experimentation.

Antisense oligonucleotides can also be a-anomeric nucleic acids, which  
15 form specific double-stranded hybrids with complementary RNA in which, contrary to the usual b-units, the strands run parallel to each other [Gaultier et al., *Nucleic Acids Res.* 15:6625-6641, (1987)]. Alternatively, antisense nucleic acids can comprise a 2'-o-methylribonucleotide [Inoue et al., *Nucleic Acids Res.* 15:6131-6148, (1987)] or a chimeric RNA-DNA analogue [Inoue et al.,  
20 *FEBS Lett.* 215:327-330, (1987)].

The nucleic acid sequences described above can also include ribozymes catalytic sequences. Such a ribozyme will have specificity for a protein encoded by the novel nucleic acids described herein (by virtue of having one or more sequences that are complementary to the cDNAs that represent novel  
25 genes or the novel portions (i.e., the portions not found in related splice variants) of the sequences that represent new splice variants. These ribozymes can include a catalytic sequence encoding a protein that cleaves mRNA [see U.S. Pat. No. 5,093,246 or Haselhoff and Gerlach, *Nature* 334:585-591, (1988)]. For example, a derivative of a tetrahymena L-19 IVS RNA can be  
30 constructed in which the nucleotide sequence of the active site is

complementary to the nucleotide sequence to be cleaved in an mRNA of the invention (e.g., one of the nucleic acid sequences of the present invention; see, U.S. Patent Nos. 4,987,071 and 5,116,742). Alternatively, the mRNA sequences of the present invention can be used to select a catalytic RNA having  
5 a specific ribonuclease activity from a pool of RNA molecules [see, e.g., Bartel and Szostak, *Science* 261:1411-1418, (1993); see also Krol et al., *Bio-Techniques* 6:958-976, (1988)].

Fragments having as few as 9-10 nucleotides (e.g., 12-14, 15-17, 18-20, 21-23, or 24-27 nucleotides) can be useful as probes or expression templates  
10 and are within the scope of the present invention. Indeed, fragments that contain about 15-20 nucleotides can be used in Southern blotting, Northern blotting, dot or slot blotting, PCR amplification methods (where naturally occurring or mutant nucleic acids are amplified), colony hybridization methods, in situ hybridization, and the like.

The present invention also encompasses pairs of oligonucleotides (these  
15 can be used, for example, to amplify the new genes, or portions thereof, or the novel portions of the splice variant in, for example, potentially diseased tissue) and groups of oligonucleotides (e.g., groups that exhibit a certain degree of homology (e.g., nucleic acids that are 90% identical to one another) or that  
20 share one or more functional attributes).

When used, for example, as probes, the nucleic acids of the invention can be labeled with a radioactive isotope (e.g., using polynucleotide kinase to add <sup>32</sup>P-labeled ATP to the oligonucleotide used as the probe) or an enzyme. Other labels, such as chemiluminescent, fluorescent, or colorimetric, labels can be  
25 used.

As noted above, the invention features nucleic acids that are complementary to those represented by the nucleic acid sequences of the present invention or novel portions thereof (i.e., novel fragments) and as such  
30 are capable of hybridizing therewith. In many cases, nucleic acids that are used

as probes or primers are absolutely or completely complementary to all, or a portion of, the target sequence. However, this is not always necessary. The sequence of a useful probe or primer can differ from that of a target sequence so long as it hybridizes with the target under the stringency conditions described  
5 herein (or the conditions routinely used to amplify sequences by PCR) to form a stable duplex.

Hybridization of a nucleic acid probe to sequences in a library or other sample of nucleic acids is typically performed under moderate to high stringency conditions. Nucleic acid duplex or hybrid stability is expressed as  
10 the melting temperature ( $T_m$ ), which is the temperature at which a probe dissociates from a target DNA and, therefore, helps define the required stringency conditions. To identify sequences that are related or substantially identical to that of a probe, it is useful to first establish the lowest temperature at which only homologous hybridization occurs with a particular concentration of  
15 salt (e.g., SSC or SSPE). (The terms "identity" or "identical" as used herein are equated with the terms "homology" or "homologous"). Then, assuming a 1% mismatch requires a 1°C decrease in the  $T_m$ , the temperature of the wash (e.g., the final wash) following the hybridization reaction is reduced accordingly. For example, if sequences having at least 95% identity with the probe are sought,  
20 the final wash temperature is decreased by 5°C. In practice, the change in  $T_m$  can be between 0.5°C and 1.5°C per 1% mismatch

The hybridization conditions described here can be employed when the nucleic acids of the invention are used in, for example, diagnostic assays, or when one wishes to identify, for example, the homologous genes that fall within  
25 the scope of the invention (as stated elsewhere, the invention encompasses allelic variants, homologues and orthologues of the sequences that represent new genes). Homologous genes will hybridize with the sequences that represent new genes under a stringency condition described herein.

A hybridization reaction is carried out at "high stringency" if  
30 hybridization (between the probe and a potential target sequence) is carried out

at 68°C in (a) 5X SSC/5X Denhardt's solution/1.0% SDS, (b) 0.5 M NaHPO<sub>4</sub> (pH 7.2)/1 mM EDTA/7% SDS, or (c) 50% formamide/0.25 M NaHPO<sub>4</sub> (pH 7.2)/0.25 M NaCl/1 mM EDTA/7% SDS, and washing is carried out with (a) 0.2X SSC/0.1% SDS at room temperature or at 42°C, (b) 0.1X SSC/0.1% SDS  
5 at 68°C, or (c) 40 mM NaHPO<sub>4</sub> (pH 7.2)/1 mM EDTA and either 1% or 5% SDS at 50°C.

"Moderately stringent" conditions constitute the hybridization conditions described above and one or more washes in 3X SSC at 42°C. Of course, salt concentration and temperature can be varied to achieve the optimal level of  
10 identity between the probe and the target nucleic acid. This is well known in the art, and additional guidance is available in, for example, Sambrook et al., 1989, Molecular Cloning, A Laboratory Manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., and Ausubel et al. (eds.), 1995, Current Protocols in Molecular Biology, John Wiley & Sons, New York, N.Y.

15 As mentioned hereinabove, the nucleic acid sequences of the present invention can be modified to encode substitution mutants of the wild type forms. Substitution mutants can include amino acid residues that represent either a conservative or non-conservative change (or, where more than one residue is varied, possibly both). A "conservative" substitution is one in which  
20 one amino acid residue is replaced with another having a similar side chain. Families of amino acid residues having similar side chains have been defined in the art. These families include amino acids with basic side chains (e.g., lysine, arginine, histidine), acidic side chains (e.g., aspartic acid, glutamic acid), uncharged polar side chains (e.g., glycine, asparagine, glutamine, serine, threonine, tyrosine, cysteine), nonpolar side chains (e.g., alanine, valine,  
25 leucine, isoleucine, proline, phenylalanine, methionine, tryptophan), beta-branched side chains (e.g., threonine, valine, isoleucine) and aromatic side chains (e.g., tyrosine, phenylalanine, tryptophan, histidine). The invention includes polypeptides that include one, two, three, five, or more conservative  
30 amino acid substitutions, where the resulting mutant polypeptide has at least

one biological activity that is the same, or substantially the same, as a biological activity of the wild type polypeptide.

Fragments or other mutant nucleic acids can be made by mutagenesis techniques well known in the art, including those applied to polynucleotides, cells, or organisms (e.g., mutations can be introduced randomly along all or part of the nucleic acid sequences of the present invention by saturation mutagenesis). The resultant mutant proteins can be screened for biological activity to identify those that retain activity or exhibit altered activity.

In certain embodiments, nucleic acids of the invention differ from the nucleic acid sequences provided in files "Transcripts\_nucleotide\_seqs\_part1", "Transcripts\_nucleotide\_seqs\_part2", "Transcripts\_nucleotide\_seqs\_part3.new", "Transcripts\_nucleotide\_seqs\_part4", and "ProDG\_seqs" (provided in CD-ROM1 and CD-ROM2) by at least one, but less than 10, 20, 30, 40, 50, 100, or 200 nucleotides or, alternatively, at less than 1%, 5%, 10% or 20% of the nucleotides in the subject nucleic acid (excluding, of course, splice variants known in the art). Similarly, in certain embodiments, proteins of the invention can differ from those encoded by those included in File "Protein.seqs" (provided in CD-ROM2) by at least one, but less than 10, 20, 30, 40, 50, 100, or 200 amino acid residues or, alternatively, at less than 1%, 5%, 10% or 20% of the amino acid residues in a subject protein (excluding, of course, proteins encoded by splice variants known in the art (proteins of the invention are described in more detail below)). If necessary for this analysis (or any other test for homology or substantial identity described herein), the sequences should be aligned for maximum homology, as described elsewhere here.

The present invention also encompasses mutants [e.g., nucleic acids that are 80% (or more) identical to one of the nucleic acid sequences disclosed in CD-ROMs 1 and 2], which encode proteins that retain substantially at least one, or preferably substantially all of the biological activities of the referenced protein. What constitutes "substantially all" may vary considerably. For example, in

some instances, a variant or mutant protein may be about 5% as effective as the protein from which it was derived. But if that level of activity is sufficient to achieve a biologically significant result (e.g., transport of a sufficient number of ions across a cell membrane), the variant or mutant protein is one that retains substantially all of at least one of the biological activities of the protein from which it was derived. A "biologically active" variant or mutant (e.g., fragment) of a protein can participate in an intra- or inter-molecular interaction that can be characterized by specific binding between molecules two or more identical molecules (in which case, homodimerization could occur) or two or more different molecules (in which case, heterodimerization could occur). Often, a biologically active fragment will be recognizable by virtue of a recognizable domain or motif, and one can confirm biological activity experimentally. More specifically, for example, one can make (by synthesis or recombinant techniques) a nucleic acid fragment that encodes a potentially biologically active portion of a protein of the present invention by inserting the active fragment into an expression vector, and expressing the protein (genetic constructs and expression systems are described further below), and finally assessing the ability of the protein to function.

The present invention also encompasses chimeric nucleic acid sequences that encode fusion proteins. For example, a nucleic acid sequence of the invention can include a sequence that encodes a hexa-histidine tag (to facilitate purification of bacterially-expressed proteins) or a hemagglutinin tag (to facilitate purification of proteins expressed in eukaryotic cells).

The fused heterologous sequence can also encode a portion of an immunoglobulin (e.g., the constant region (Fc) of an IgG molecule), a detectable marker, or a signal sequence (e.g., a sequence that is recognized and cleaved by a signal peptidase in the host cell in which the fusion protein is expressed). Fusion proteins containing an Fc region can be purified using a protein A column, and they have increased stability (e.g., a greater circulating half-life) in vivo.

Detectable markers are well known in the art and can be used in the context of the present invention. For example, the expression vector pUR278 (Ruther et al., EMBO J., 2:1791, 1983) can be used to fuse a nucleic acid of the invention to the lacZ gene (which encodes  $\beta$ -galactosidase).

5 A nucleic acid sequence of the invention can also be fused to a sequence that, when expressed, improves the quantity or quality (e.g., solubility) of the fusion protein. For example, pGEX vectors can be used to express the proteins of the invention fused to glutathione S-transferase (GST). In general, such fusion proteins are soluble and can be easily purified from lysed cells by  
10 adsorption to glutathione-agarose beads followed by elution in the presence of free glutathione. The pGEX vectors (Pharmacia Biotech Inc; Smith and Johnson, Gene 67:31-40, 1988) are designed to include thrombin or factor Xa protease cleavage sites so that the cloned target gene product can be released from the GST moiety. Other useful vectors include pMAL (New England  
15 Biolabs, Beverly, MA) and pRIT5 (Pharmacia, Piscataway, NJ), which fuse maltose E binding protein and protein A, respectively, to a protein of the invention.

A signal sequence, when present, can facilitate secretion of the fusion protein from a cell, and can be cleaved off by the host cell. The nucleic acid  
20 sequences of the present invention can also be fused to "inactivating" sequences, which render the fusion protein encoded, as a whole, inactive. Such proteins can be referred to as "preproteins," and they can be converted into an active form of the protein by removal of the inactivating sequence.

The present invention also encompasses genetic constructs (e.g.,  
25 plasmids, cosmids, and other vectors that transport nucleic acids) that include a nucleic acid of the invention in a sense or antisense orientation. The nucleic acids can be operably linked to a regulatory sequence (e.g., a promoter, enhancer, or other expression control sequence, such as a polyadenylation signal) that facilitates expression of the nucleic acid. The vector can replicate



autonomously or integrate into a host genome, and can be a viral vector, such as a replication defective retrovirus, an adenovirus, or an adeno-associated virus.

When present, the regulatory sequence can direct constitutive or tissue-specific expression of the nucleic acid. Tissue-specific promoters include, for example, the liver-specific albumin promoter (Pinkert et al., *Genes Dev.* 1:268-277, 1987), lymphoid-specific promoters (Calame and Eaton, *Adv. Immunol.* 43:235-275, 1988), such as those of T cell receptors (Winoto and Baltimore, *EMBO J.* 8:729-733, 1989) and immunoglobulins (Banerji et al., *Cell* 33:729-740, 1982; Queen and Baltimore, *Cell* 33:741-748, 1983), the neuron-specific neurofilament promoter (Byrne and Ruddle, *Proc. Natl. Acad. Sci. USA* 86:5473-5477, 1989), pancreas-specific promoters (Edlund et al., *Science* 230:912-916, 1985), and mammary gland-specific promoters (e.g., milk whey promoter; see U.S. Patent No. 4,873,316 and European Application Publication No. 264,166). Developmentally-regulated promoters can also be used. Examples of such promoters include the murine hox promoters (Kessel and Gruss, *Science* 249:374-379, 1990) and the fetoprotein promoter (Campes and Tilghman, *Genes Dev.* 3:537-546, 1989). Moreover, the promoter can be an inducible promoter. For example, the promoter can be regulated by a steroid hormone, a polypeptide hormone, or some other polypeptide (e.g., that used in the tetracycline-inducible system, "Tet-On" and "Tet-Off"; see, e.g., Clontech Inc. (Palo Alto, CA), Gossen and Bujard *Proc. Natl. Acad. Sci. USA* 89:5547, 1992, and Paillard, *Human Gene Therapy* 9:983, 1989).

The expression vector will be selected or designed depending on, for example, the type of host cell to be transformed and the level of protein expression desired. For example, when the host cells are mammalian cells, the expression vector can include viral regulatory elements, such as promoters derived from polyoma, Adenovirus 2, cytomegalovirus and Simian Virus 40. The nucleic acid inserted (i.e., the sequence to be expressed) can also be modified to encode residues that are preferentially utilized in *E. coli* (Wada et

al., Nucleic Acids Res. 20:2111-2118, 1992). These modifications can be achieved by standard DNA synthesis techniques.

Expression vectors can be used to produce the proteins encoded by the nucleic acid sequences of the invention *ex vivo* (e.g., the expressed proteins can be purified from expression systems such as those described herein) or *in vivo* (in, for example, whole organisms). Proteins can be expressed *in vivo* in a way that restores expression to within normal limits and/or restores the temporal or spatial patterns of expression normally observed. Alternatively, proteins can be aberrantly expressed *in vivo* (i.e., at a time or place, or to an extent, that does not normally occur *in vivo*). For example, proteins can be over expressed or under expressed with respect to expression in a wild-type state; expressed at a different developmental stage; expressed at a different time during the cell cycle; or expressed in a tissue or cell type where expression does not normally occur.

The present invention also encompasses various engineered cells, including cells that have been engineered to express or over-express a nucleic acid sequence described herein. Accordingly, the cells can be transformed with a genetic construct, such as those described above. A "transformed" cell is a cell into which (or into an ancestor of which) one has introduced a nucleic acid that encodes a protein of the invention. The nucleic acid can be introduced by any of the art-recognized techniques for introducing nucleic acids into a host cell (e.g., calcium phosphate or calcium chloride co-precipitation, DEAE-dextran-mediated transfection, lipofection, or electroporation).

The terms "transformed cell" or "host cell" refer not only to the particular subject cell, but also to the progeny or potential progeny of such cells. Mutations or environmental influences may modify the cells in succeeding generations and, even though such progeny may not be identical to the parent cell, they are nevertheless within the scope of the invention. The cells of the invention can be "isolated" cells or "purified preparations" of cells (e.g., an *in vitro* preparation of cells), either of which can be obtained from multicellular

organisms such as plants and animals (in which case the purified preparation would constitute a subset of the cells from the organism). In the case of unicellular microorganisms (e.g., microbial cells), the preparation is purified when at least 10% (e.g., 25%, 50%, 75%, 80%, 90%, 95% or more) of the cells within it are the cells of interest (e.g., the cells that express a protein of the invention).

The expression vectors of the invention can be designed to express proteins in prokaryotic or eukaryotic cells. For example, polypeptides of the invention can be expressed in bacterial cells (e.g., *E. coli*), fungi, yeast, or insect cells (e.g., using baculovirus expression vectors). For example, a baculovirus such as *Autographa californica* nuclear polyhedrosis virus (AcNPV), which grows in *Spodoptera frugiperda* cells, can be used as a vector to express foreign genes. A nucleic acid of the invention can be cloned into a non-essential region (for example the polyhedrin gene) of the viral genome and placed under control of a promoter (e.g., the polyhedrin promoter). Successful insertion of the nucleic acid results in inactivation of the polyhedrin gene and production of non-occluded recombinant virus (i.e., virus lacking the proteinaceous coat encoded by the polyhedrin gene). These recombinant viruses are then typically used to infect insect cells (e.g., *Spodoptera frugiperda* cells) in which the inserted gene is expressed (see, e.g., Smith et al., *J. Virol.* 46:584, 1983 and U.S. Patent No. 4,215,051). If desired, mammalian cells can be used in lieu of insect cells, provided the virus is engineered so that the nucleic acid is placed under the control of a promoter that is active in mammalian cells.

Useful mammalian cells include rodent cells, such as Chinese hamster ovary cells (CHO) or COS cells, primate cells, such as African green monkey kidney cells, rabbit cells, or pig cells). The mammalian cells can also be human cells (e.g., a hematopoietic cell, a fibroblast, or a tumor cell). For example, HeLa cells, 293 cells, 3T3 cells, and WI38 cells are useful. Other suitable host cells are known to those skilled in the art and are discussed further in Goeddel

[Gene Expression Technology: Methods in Enzymology 185, Academic Press, San Diego, CA, (1990)].

Proteins can also be produced in plant cells, if desired. For plant cells, viral expression vectors (e.g., cauliflower mosaic virus and tobacco mosaic virus) and plasmid expression vectors (e.g., Ti plasmid) are suitable. These cells and other types are available from a wide range of sources [e.g., the American Type Culture Collection, Manassas, VA; see also, e.g., Ausubel et al., Current Protocols in Molecular Biology, John Wiley & Sons, New York, (1994)]. The optimal methods of transformation (by, for example, transfection) and, as noted above, the choice of expression vehicle will depend on the host system selected. Transformation and transfection methods are described in, for example, Ausubel et al., supra; expression vehicles can be chosen from those provided in, for example, Pouwels et al., Cloning Vectors: A Laboratory Manual, (1985), Supp. (1987). The host cells harboring the expression vehicle can be cultured in conventional nutrient media, adapted as needed for activation of a chosen nucleic acid, repression of a chosen nucleic acid, selection of transformants, or amplification of a chosen nucleic acid.

Expression systems can be selected based on their ability to produce proteins that are modified (e.g., by phosphorylation, glycosylation, or cleavage) in substantially the same way they would be in a cell in which they are naturally expressed. Alternatively, the system can be one in which naturally occurring modifications do not occur, or occur in a different position, or to a different extent, than they otherwise would.

If desired, the host cells can be those of a stably-transfected cell line. Vectors suitable for stable transfection of mammalian cells are available to the public (see, e.g., Pouwels et al. (supra) as are methods for constructing them (see, e.g., Ausubel et al. (supra). In one example, a nucleic acid of the invention is cloned into an expression vector that includes the dihydrofolate reductase (DHFR) gene. Integration of the plasmid and, therefore, the nucleic acid it

contains, into the host cell chromosome is selected for by including 0.01-300 mM methotrexate in the cell culture medium (as described in Ausubel et al., supra). This dominant selection can be accomplished in most cell types.

Moreover, recombinant protein expression can be increased by DHFR-mediated amplification of the transfected gene. Methods for selecting cell lines bearing gene amplifications are described in Ausubel et al. (supra) and generally involve extended culture in medium containing gradually increasing levels of methotrexate. DHFR-containing expression vectors commonly used for this purpose include pCVSEII-DHFR and pAdD26SV(A) (which are also described in Ausubel et al., supra).

A number of other selection systems can be used. These include those based on herpes simplex virus thymidine kinase, hypoxanthine-guanine phosphoribosyl-transferase, and adenine phosphoribosyltransferase genes, which can be employed in tk, hgprt, or aprt cells, respectively. In addition, gpt, which confers resistance to mycophenolic acid (Mulligan et al., Proc. Natl. Acad. Sci. USA, 78:2072, 1981); neo, which confers resistance to the aminoglycoside G-418 (Colberre-Garapin et al., J. Mol. Biol. 150:1, 1981); and hygromycin, which confers resistance to hygromycin (Santerre et al., Gene 30:147, 1981), can be used.

In view of the foregoing, it is clear that one can synthesize proteins encoded by the nucleic acid sequences of the present invention (i.e., recombinant proteins). Methods of generating and recombinant proteins are well known in the art. Recombinant protein purification can be effected by affinity. Where a protein of the invention has been fused to a heterologous protein (e.g., a maltose binding protein, a  $\beta$ -galactosidase protein, or a trpE protein), antibodies or other agents that specifically bind to the latter can facilitate purification. The recombinant protein can, if desired, be further purified (e.g., by high performance liquid chromatography or other standard techniques [see, Fisher, Laboratory Techniques In Biochemistry And Molecular Biology, Eds., Work and Burdon, Elsevier, (1980)]).

Other purification schemes are known as well. For example, non-denatured fusion proteins can be purified from human cell lines as described by Janknecht et al. (Proc. Natl. Acad. Sci. USA, 88:8972, 1981). In this system, a nucleic acid is subcloned into a vaccinia recombination plasmid such that it is translated, in frame, with a sequence encoding an N-terminal tag consisting of six histidine residues. Extracts of cells infected with the recombinant vaccinia virus are loaded onto  $\text{Ni}^{2+}$  nitriloacetic acid-agarose columns, and histidine-tagged proteins are selectively eluted with imidazole-containing buffers.

Alternatively, Chemical synthesis can also be utilized to generate the proteins of the present invention [e.g., proteins can be synthesized by the methods described in Solid Phase Peptide Synthesis, 2nd Ed., The Pierce Chemical Co., Rockford, IL, (1984)].

The invention also features expression vectors that can be transcribed and translated in vitro using, for example, a T7 promoter and T7 polymerase. Thus, the invention encompasses methods of making the proteins described herein in vitro.

Sufficiently purified proteins can be used as described herein. For example, one can administer the protein to a patient, use it in diagnostic or screening assays, or use it to generate antibodies (these methods are described further below).

The cells per se can also be administered to patients in the context of replacement therapies. For example, a nucleic acid of the present invention can be operably linked to an inducible promoter (e.g., a steroid hormone receptor-regulated promoter) and introduced into a human or nonhuman (e.g., porcine) cell and then into a patient. Optionally, the cell can be cultivated for a time or encapsulated in a biocompatible material, such as poly-lysine alginate. See, e.g., Lanza, Nature Biotechnol. 14:1107, (1996); Joki et al. Nature Biotechnol. 19:35, 2001; and U.S. Patent No. 5,876,742] When a steroid hormone receptor-regulated promoter is used, protein production can be regulated in the subject by administering a steroid hormone to the subject. Implanted recombinant cells

can also express and secrete an antibody that specifically binds to one of the proteins encoded by the nucleic acid sequences of the present invention. The antibody can be any antibody or any antibody derivative described herein. An antibody "specifically binds" to a particular antigen when it binds to that antigen but not, to a detectable level, to other molecules in a sample (e.g., a tissue or cell culture) that naturally includes the antigen.

While the host cells described above express recombinant proteins, the invention also encompasses cells in which gene expression is disrupted (e.g., cells in which a gene has been knocked out). These cells can serve as models of disorders that are related to mutated or mis-expressed alleles and are also useful in drug screening.

Protein expression can also be regulated in cells without using the genetic constructs described above. Instead, one can modify the expression of an endogenous gene within a cell (e.g., a cell line or microorganism) by inserting a heterologous DNA regulatory element into the genome of the cell such that the element is operably linked to the endogenous gene. For example, an endogenous gene that is "transcriptionally silent," (i.e., not expressed at detectable levels) can be activated by inserting a regulatory element that promotes the expression of a normally expressed gene product in that cell. Techniques such as targeted homologous recombination can be used to insert the heterologous DNA (see, e.g., U.S. Patent No. 5,272,071 and WO 91/06667).

The polypeptides of the present invention include the protein sequences contained in the File "Protein.seqs" of CD-ROM 2 and those encoded by the nucleic acids described herein (so long as those nucleic acids contain coding sequence and are not wholly limited to an untranslated region of a nucleic acid sequence), regardless of whether they are recombinantly produced (e.g., produced in and isolated from cultured cells), otherwise manufactured (by, for example, chemical synthesis), or isolated from a natural biological source (e.g., a cell or tissue) using standard protein purification techniques.

The terms "peptide," "polypeptide," and "protein" are used herein interchangeably to refer to a chain of amino acid residues, regardless of length or post-translational modification (e.g., glycosylation or phosphorylation). Proteins (including antibodies that specifically bind to the products of those nucleic acid sequences that encode protein or fragments thereof) and other compounds can be "isolated" or "purified." The proteins and compounds of the present invention are "isolated" or "purified" when they exist as a composition that is at least 60% (e.g., 70%, 75%, 80%, 85%, 90%, 95%, or 99% or more) by weight the protein or compound of interest. Thus, the proteins of the invention are substantially free from the cellular material (or other biological or cell culture material) with which they may have, at one time, been associated (naturally or otherwise). Purity can be measured by any appropriate standard method (e.g., column chromatography, polyacrylamide gel electrophoresis, or HPLC analysis

The proteins of the invention also include those encoded by novel fragments or other mutants or variants of the protein-encoding sequences of the present invention. These proteins can retain substantially all (e.g., 70%, 80%, 90%, 95%, or 99%) of the biological activity of the full-length protein from which they were derived and can, therefore, be used as agonists or mimetics of the proteins from which they were derived. The manner in which biological activity can be determined is described generally herein, and specific assays (e.g., assays of enzymatic activity or ligand-binding ability) are known to those of ordinary skill in the art. In some instances, retention of biological activity is not necessary or desirable. For example, fragments that retain little, if any, of the biological activity of a full-length protein can be used as immunogens, which, in turn, can be used as therapeutic agents (e.g., to generate an immune response in a patient), diagnostic agents (e.g., to detect the presence of antibodies or other proteins in a tissue sample obtained from a patient), or to generate or test antibodies that specifically bind the proteins of the invention.



In other instances, the proteins encoded by nucleic acids of the invention can be modified (e.g., fragmented or otherwise mutated) so their activities oppose those of the naturally occurring protein (i.e., the invention encompasses variants of the proteins encoded by nucleic acids of the invention that are antagonistic to a biological process). One of ordinary skill in the art will recognize that the more extensive the mutation, the more likely it is to affect the biological activity of the protein (this is not to say that minor modifications cannot do so as well). Thus, it is likely that mutant proteins that are agonists of those encoded by wild type proteins will differ from those wild type proteins only at non-essential residues or will contain only conservative substitutions. Conversely, antagonists are likely to differ at an essential residue or to contain non-conservative substitutions. Moreover, those of ordinary skill in the art can engineer proteins so that they retain desirable traits (i.e., those that make them efficacious in a particular therapeutic, diagnostic, or screening regime) and lose undesirable traits (i.e., those that produce side effects, or produce false-positive results through non-specific binding).

In the event a protein of the invention is encoded by a new gene, the invention encompasses proteins that arise following alternative transcription, RNA splicing, translational- or post-translational events (e.g., the invention encompasses splice variants of the new genes). In the event a protein of the invention is encoded by a novel splice variant, the invention encompasses proteins that arise following alternative translational- or post-translational events (i.e., the invention does not encompass proteins encoded by known splice variants, but does encompass other variants of the novel splice variant). Post-translational modifications are discussed above in the context of expression systems.

The fragmented or otherwise mutant proteins of the invention can differ from those encoded by the nucleic acids of the invention to a limited extent (e.g., by at least one but less than 5, 10 or 15 amino acid residues). As with other, more extensive mutations, the differences can be introduced by adding,

deleting, and/or substituting one or more amino acid residues. Alternatively, the mutant proteins can differ from the wild type proteins from which they were derived by at least one residue but less than 5%, 10%, 15% or 20% of the residues when analyzed as described herein. If the mutant and wild type  
5 proteins are different lengths, they can be aligned and analyzed using the algorithms described above.

Useful variants, fragments, and other mutants of the proteins encoded by the nucleic acids of the invention can be identified by screening combinatorial libraries of these variants, fragments, and other mutants for agonist or  
10 antagonist activity. For example, libraries of fragments (e.g., N-terminal, C-terminal, or internal fragments) of one or more of the proteins of the invention can be used to generate populations of fragments that can be screened and, once identified, isolated. The proteins can include those in which one or more cysteine residues are added or deleted, or in which a glycosylated residue  
15 is added or deleted. Methods for screening libraries (e.g., combinatorial libraries of proteins made from point mutants or cDNA libraries) for proteins or genes having a particular property are known in the art. These methods can be adapted for rapid screening. Recursive ensemble mutagenesis (REM), a new technique that enhances the frequency of functional mutants in libraries, can be  
20 used in combination with screening assays to identify useful variants of the proteins of the present invention [Arkin and Yourvan, Proc. Natl. Acad. Sci. USA 89:7811-7815, (1992); Delgrave et al., Protein Engineering 6:327-331, (1993)].

Cell-based assays can be exploited to analyze variegated libraries  
25 constructed from one or more of the proteins of the invention. For example, cells in a cell line (e.g., a cell line that ordinarily responds to the protein(s) of interest in a substrate-dependent manner) can be transfected with a library of expression vectors. The transfected cells are then contacted with the protein and the effect of the expression of the mutant on signaling by the protein  
30 (substrate) can be detected (e.g., by measuring redox activity or protein

folding). Plasmid DNA can then be recovered from the cells that score for inhibition, or alternatively, potentiation of signaling by the protein (substrate). Individual clones are then further characterized.

The invention also contemplates antibodies (i.e., immunoglobulin molecules) that specifically bind (see the definition above) to the proteins described herein and antibody fragments (e.g., antigen-binding fragments or other immunologically active portions of the antibody). Antibodies are proteins, and those of the invention can have at least one or two heavy chain variable regions (VH), and at least one or two light chain variable regions (VL). The VH and VL regions can be further subdivided into regions of hypervariability, termed "complementarity determining regions" (CDR), which are interspersed with more highly conserved "framework regions" (FR). These regions have been precisely defined [see, Kabat et al., Sequences of Proteins of Immunological Interest, Fifth Edition, U.S. Department of Health and Human Services, NIH Publication No. 91-3242, (1991) and Chothia et al., J. Mol. Biol. 196:901-917, (1987)], and antibodies or antibody fragments containing one or more of them are within the scope of the invention.

The antibodies of the invention can also include a heavy and/or light chain constant region [constant regions typically mediate binding between the antibody and host tissues or factors, including effector cells of the immune system and the first component (C1q) of the classical complement system], and can therefore form heavy and light immunoglobulin chains, respectively. For example, the antibody can be a tetramer (two heavy and two light immunoglobulin chains, which can be connected by, for example, disulfide bonds). The heavy chain constant region contains three domains (CH1, CH2 and CH3), whereas the light chain constant region has one (CL).

An antigen-binding fragment of the invention can be: (i) a Fab fragment (i.e., a monovalent fragment consisting of the VL, VH, CL and CH1 domains); (ii) a F(ab')<sub>2</sub> fragment (i.e., a bivalent fragment containing two Fab fragments linked by a disulfide bond at the hinge region); (iii) a Fd fragment consisting of

the VH and CH1 domains; (iv) a Fv fragment consisting of the VL and VH domains of a single arm of an antibody, (v) a dAb fragment [Ward et al., Nature 341:544-546, (1989)], which consists of a VH domain; and (vi) an isolated complementarity determining region (CDR).

5 F(ab')<sub>2</sub> fragments can be produced by pepsin digestion of the antibody molecule, and Fab fragments can be generated by reducing the disulfide bridges of F(ab')<sub>2</sub> fragments. Alternatively, Fab expression libraries can be constructed [Huse et al., Science 246:1275, (1989)] to allow rapid and easy identification of monoclonal Fab fragments with the desired specificity. Methods of making  
10 other antibodies and antibody fragments are known in the art. For example, although the two domains of the Fv fragment, VL and VH, are coded for by separate genes, they can be joined, using recombinant methods or a synthetic linker that enables them to be made as a single protein chain in which the VL and VH regions pair to form monovalent molecules [known as single chain Fv (scFv); see e.g., Bird et al., Science 242:423-426, (1988); Huston et al., Proc. Natl. Acad. Sci. USA 85:5879-5883, (1988); Colcher et al., Ann. NY Acad. Sci. 880:263-80, (1999); and Reiter, Clin. Cancer Res. 2:245-52, (1996)].  
15 Techniques for producing single chain antibodies are also described in U.S. Patent Nos. 4,946,778 and 4,704,692. Such single chain antibodies are encompassed within the term "antigen-binding fragment" of an antibody. These antibody fragments are obtained using conventional techniques known to those of ordinary skill in the art, and the fragments are screened for utility in the same manner that intact antibodies are screened. Moreover, a single chain antibody can form dimers or multimers and, thereby, become a multivalent antibody  
20 having specificities for different epitopes of the same target protein.

The antibody can be a polyclonal (i.e., part of a heterogeneous population of antibody molecules derived from the sera of the immunized animals) or a monoclonal antibody (i.e., part of a homogeneous population of antibodies to a particular antigen), either of which can be recombinantly  
30 produced (e.g., produced by phage display or by combinatorial methods, as

described in, e.g., U.S. Patent No. 5,223,409; WO 92/18619; WO 91/17271; WO 92/20791; WO 92/15679; WO 93/01288; WO 92/01047; WO 92/09690; WO 90/02809; Fuchs et al., *Bio/Technology* 9:1370-1372, (1991); Hay et al. *Human Antibody Hybridomas* 3:81-85, (1992); Huse et al. *Science* 246:1275-1281, (1989); Griffiths et al. *EMBO J* 12:725-734, (1993); Hawkins et al., *J. Mol Biol* 226:889-896, (1992); Clackson et al. *Nature* 352:624-628, (1991); Gram et al., *Proc. Natl. Acad. Sci. USA* 89:3576-3580, (1992); Garrad et al., *Bio/Technology* 9:1373-1377, (1991); Hoogenboom et al. *Nucl. Acids Res.* 19:4133-4137, (1991); and Barbas et al., *Proc. Natl. Acad. Sci. USA* 88:7978-7982, (1991). In one embodiment, an antibody is made by immunizing an animal with a protein encoded by a nucleic acid of the invention (one, of course, that contains coding sequence) or a mutant or fragment (e.g., an antigenic peptide fragment) thereof. Alternatively, an animal can be immunized with a tissue sample (e.g., a crude tissue preparation, a whole cell (living, lysed, or fractionated) or a membrane fraction). Thus, antibodies of the invention can specifically bind to a purified antigen or a tissue (e.g., a tissue section, a whole cell (living, lysed, or fractionated) or a membrane fraction).

In the event an antigenic peptide is used, it can include at least eight (e.g., 10, 15, 20, or 30) consecutive amino acid residues found in a protein of the invention. The antibodies generated can specifically bind to one of the proteins in their native form (thus, antibodies with linear or conformational epitopes are within the invention), in a denatured or otherwise non-native form, or both. Conformational epitopes can sometimes be identified by identifying antibodies that bind to a protein in its native form, but not in a denatured form.

The host animal (e.g., a rabbit, mouse, guinea pig, or rat) can be immunized with the antigen, optionally linked to a carrier (i.e., a substance that stabilizes or otherwise improves the immunogenicity of an associated molecule), and optionally administered with an adjuvant (see, e.g., Ausubel et al., *supra*). An exemplary carrier is keyhole limpet hemocyanin (KLH) and exemplary adjuvants, which will be selected in view of the host animal's

species, include Freund's adjuvant (complete or incomplete), adjuvant mineral gels (e.g., aluminum hydroxide), surface active substances such as lysolecithin, pluronic polyols, polyanions, peptides, oil emulsions, dinitrophenol, BCG (bacille Calmette-Guerin), and *Corynebacterium parvum*. KLH is also  
5 sometimes referred to as an adjuvant. The antibodies generated in the host can be purified by, for example, affinity chromatography methods in which the polypeptide antigen is immobilized on a resin.

Epitopes encompassed by an antigenic peptide may be located on the surface of the protein (e.g., in hydrophilic regions), or in regions that are highly  
10 antigenic (such regions can be selected, initially, by virtue of containing many charged residues). An Emini surface probability analysis of human protein sequences can be used to indicate the regions that have a particularly high probability of being localized to the surface of the protein.

The antibody can be a fully human antibody (e.g., an antibody made in a  
15 mouse that has been genetically engineered to produce an antibody from a human immunoglobulin sequence, such as that of a human immunoglobulin gene (the kappa, lambda, alpha (IgA1 and IgA2), gamma (IgG1, IgG2, IgG3, IgG4), delta, epsilon and mu constant region genes or the myriad immunoglobulin variable region genes). Alternatively, the antibody can be a  
20 non-human antibody (e.g., a rodent (e.g., a mouse or rat), goat, or non-human primate (e.g., monkey) antibody).

Methods of producing antibodies are well known in the art. For example, as noted above, human monoclonal antibodies can be generated in transgenic mice carrying the human immunoglobulin genes rather than those of  
25 the mouse. Splenocytes obtained from these mice (after immunization with an antigen of interest) can be used to produce hybridomas that secrete human mAbs with specific affinities for epitopes from a human protein (see, e.g., WO 91/00906, WO 91/10741; WO 92/03918; WO 92/03917; Lonberg et al., Nature 368:856-859, 1994; Green et al., Nature Genet. 7:13-21, 1994; Morrison et al.  
30 Proc. Natl. Acad. Sci. USA 81:6851-6855, 1994; Bruggeman et al., Immunol.

7:33-40, 1993; Tuaillon et al., Proc. Natl. Acad. Sci. USA 90:3720-3724, 1993; and Bruggeman et al., Eur. J. Immunol 21:1323-1326, 1991).

The antibody can also be one in which the variable region, or a portion thereof (e.g., a CDR), is generated in a non-human organism (e.g., a rat or mouse). Thus, the invention encompasses chimeric, CDR-grafted, and humanized antibodies and antibodies that are generated in a non-human organism and then modified (in, e.g., the variable framework or constant region) to decrease antigenicity in a human. Chimeric antibodies (i.e., antibodies in which different portions are derived from different animal species (e.g., the variable region of a murine mAb and the constant region of a human immunoglobulin) can be produced by recombinant techniques known in the art. For example, a gene encoding the Fc constant region of a murine (or other species) monoclonal antibody molecule can be digested with restriction enzymes to remove the region encoding the murine Fc, and the equivalent portion of a gene encoding a human Fc constant region can be substituted therefore [see European Patent Application Nos. 125,023; 184,187; 171,496; and 173,494; see also WO 86/01533; U.S. Patent No. 4,816,567; Better et al., Science 240:1041-1043, (1988); Liu et al., Proc. Natl. Acad. Sci. USA 84:3439-3443, (1987); Liu et al., J. Immunol. 139:3521-3526, (1987); Sun et al., Proc. Natl. Acad. Sci. USA 84:214-218, (1987); Nishimura et al., Cancer Res. 47:999-1005, (1987); Wood et al., Nature 314:446-449, (1985); Shaw et al., J. Natl. Cancer Inst. 80:1553-1559, (1988); Morrison et al., Proc. Natl. Acad. Sci. USA 81:6851, (1984); Neuberger et al., Nature 312:604, (1984); and Takeda et al., Nature 314:452, (1984)].

In a humanized or CDR-grafted antibody, at least one or two, but generally all three of the recipient CDRs (of heavy and or light immunoglobulin chains) will be replaced with a donor CDR. One need only replace the number of CDRs required for binding of the humanized antibody to a protein described herein or a fragment thereof. The donor can be a rodent antibody, and the recipient can be a human framework or a human consensus framework.

Typically, the immunoglobulin providing the CDRs is called the "donor" (and is often that of a rodent) and the immunoglobulin providing the framework is called the "acceptor." The acceptor framework can be a naturally occurring (e.g., a human) framework, a consensus framework or sequence, or a sequence  
5 that is at least 85% (e.g., 90%, 95%, 99%) identical thereto. A "consensus sequence" is one formed from the most frequently occurring amino acids (or nucleotides) in a family of related sequences (see, e.g., Winnaker, *From Genes to Clones*, Verlagsgesellschaft, Weinheim, Germany, 1987). Each position in the consensus sequence is occupied by the amino acid residue that occurs most  
10 frequently at that position in the family (where two occur equally frequently, either can be included). A "consensus framework" refers to the framework region in the consensus immunoglobulin sequence.

An antibody can be humanized by methods known in the art. For example, humanized antibodies can be generated by replacing sequences of the  
15 Fv variable region that are not directly involved in antigen binding with equivalent sequences from human Fv variable regions. General methods for generating humanized antibodies are provided by Morrison [Science 229:1202-1207, (1985)], Oi et al. [BioTechniques 4:214, (1986)], and Queen et al. (US Patent Nos. 5,585,089; 5,693,761 and 5,693,762). Those nucleic acid  
20 sequences required by these methods can be obtained from a hybridoma producing an antibody the polypeptides of the present invention, or fragments thereof. The recombinant DNA encoding the humanized antibody, or fragment thereof, can then be cloned into an appropriate expression vector.

Humanized or CDR-grafted antibodies can be produced such that one,  
25 two, or all CDRs of an immunoglobulin chain can be replaced [see, e.g., U.S. Patent No. 5,225,539; Jones et al., *Nature* 321:552-525, (1986); Verhoeyan et al., *Science* 239:1534, (1988); and Beidler et al., *J. Immunol.* 141:4053-4060, (1988)]. Thus, the invention features humanized antibodies in which specific amino acid residues have been substituted, deleted or added (in, e.g., in the  
30 framework region to improve antigen binding). For example, a humanized



antibody will have framework residues identical to those of the donor or to amino acid residues other than those of the recipient framework residue. To generate such antibodies, a selected, small number of acceptor framework residues of the humanized immunoglobulin chain are replaced by the  
5. corresponding donor amino acids. The substitutions can occur adjacent to the CDR or in regions that interact with a CDR (U.S. Patent No. 5,585,089, see especially columns 12-16). Other techniques for humanizing antibodies are described in EP 519596 A1.

In certain embodiments, the antibody has an effector function and can fix  
10 complement, while in others it can neither recruit effector cells nor fix complement. The antibody can also have little or no ability to bind an Fc receptor. For example, it can be an isotype or subtype, or a fragment or other mutant that cannot bind to an Fc receptor (e.g., the antibody can have a mutant (e.g., a deleted) Fc receptor binding region). The antibody may or may not alter  
15 (e.g., increase or decrease) the activity of a protein to which it binds.

In other embodiments, the antibody can be coupled to a heterologous substance, such as a toxin (e.g., ricin, diphtheria toxin, or active fragments thereof), another type of therapeutic agent (e.g., an antibiotic), or a detectable label. A detectable label can include an enzyme (e.g., horseradish peroxidase,  
20 alkaline phosphatase,  $\beta$ -galactosidase, or acetylcholinesterase), a prosthetic group (e.g., streptavidin/biotin and avidin/biotin), or a fluorescent, luminescent, bioluminescent, or radioactive material (e.g., umbelliferone, fluorescein, fluorescein isothiocyanate, rhodamine, dichlorotriazinylamine fluorescein, dansyl chloride or phycoerythrin (which are fluorescent), luminol (which is  
25 luminescent), luciferase, luciferin, and aequorin (which are bioluminescent), and  $^{125}\text{I}$ ,  $^{131}\text{I}$ ,  $^{35}\text{S}$  or  $^3\text{H}$  (which are radioactive)).

The antibodies of the invention (e.g., a monoclonal antibody) can be used to isolate the proteins of the invention (by, for example, affinity chromatography or immunoprecipitation) or to detect them in, for example, a  
30 cell lysate or supernatant (by Western blotting, ELISAs, radioimmune assays,

and the like) or a histological section. One can therefore determine the abundance and pattern of expression of a particular protein. This information can be useful in making a diagnosis or in evaluating the efficacy of a clinical test.

5           The invention also includes the nucleic acids that encode the antibodies described above and vectors and cells (e.g., mammalian cells such as CHO cells or lymphatic cells) that contain them. Similarly, the invention includes cell lines (e.g., hybridomas) that make the antibodies of the invention and methods of making those cell lines.

10           Non-human transgenic animals are also within the scope of the invention. These animals can be used to study the function or activity of proteins of the invention and to identify or evaluate agents that modulate their activity. A "transgenic animal" can be a mammal (e.g., a mouse, rat, dog, pig, cow, sheep, goat, or non-human primate), an avian (e.g., a chicken), or an  
15   amphibian (e.g. a frog) having one or more cells that include a transgene (e.g., an exogenous DNA molecule or a rearrangement (e.g., deletion of) endogenous chromosomal DNA). The transgene can be integrated into or can occur within the genome of the cells of the animal, and it can direct the expression of an encoded gene product in one or more types of cells or tissues. Alternatively, a  
20   transgene can "knock out" or reduce gene expression. This can occur when an endogenous gene has been altered by homologous recombination, which occurs between it and an exogenous DNA molecule that was introduced into a cell of the animal (e.g., an embryonic cell) at a very early stage in the animal's development.

25           Intronic sequences and polyadenylation signals can be included in the transgene and, when present, can increase expression. One or more tissue-specific regulatory sequences can also be operably linked to a transgene of the invention to direct expression of protein to particular cells (exemplary regulatory sequences are described above, and many others are known to those  
30   of ordinary skill in the art).

A "founder" animal is one that carries a transgene of the invention in its genome or expresses mRNA from the transgene in its cells or tissues. Founders can be bred to produce a line of transgenic animals carrying the founder's transgene or bred with founders carrying other transgenes (in which case the progeny would bear the transgenes borne by both founders). Accordingly, the invention features founder animals, their progeny, cells or populations of cells obtained therefrom, and proteins obtained therefrom. For example, a nucleic acid of the invention can be placed under the control of a promoter that directs expression of the encoded protein in the milk or eggs of the transgenic animal. The protein can then be purified or recovered from the animal's milk or eggs. Animals suitable for such purpose include pigs, cows, goats, sheep, and chickens.

The biomolecular sequences of the present invention can be divided to functional groups, according to GO classification ([www.geneontology.org](http://www.geneontology.org)), defined by the activity of the original sequences from which the new variants have been identified or to which the novel genes are homologous. Based on this classification it is possible to identify diseases and conditions which can be diagnosed and treated using novel sequence information and annotations such as those uncovered by the present invention.

***Immunoglobulin:***

This category contains proteins that are involved in the immune and complement systems such as antigens and autoantigens, immunoglobulins, MHC and HLA proteins and their associated proteins.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases involving the immunological system including inflammation, autoimmune diseases, infectious diseases, as well as cancerous processes; while probe sequences or antibodies may be used for diagnosis of such diseases.

***Transcription factor binding:***

This category contains proteins involved in transcription factors binding, RNA and DNA binding, such as transcription factors, RNA and DNA binding proteins, zinc fingers, helicase, isomerase, histones, nucleases.

5       Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases involving transcription factors binding proteins, for example diseases where there is non-normal replication or transcription of DNA and RNA respectively; while probe  
10       sequences or antibodies may be used for diagnosis of such diseases.

***Small GTPase regulatory/interacting protein:***

This category contains proteins such as RAB escort protein, guanyl-nucleotide exchange factor, guanyl-nucleotide exchange factor adaptor, GDP-dissociation inhibitor, GTPase inhibitor, GTPase activator, guanyl-nucleotide  
15       releasing factor, GDP-dissociation stimulator, regulator of G-protein signaling, RAS interactor, RHO interactor, RAB interactor, RAL interactor.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases in which the  
20       signal-transduction, typically involving G-proteases is non-normal, either as a cause, or as a result of the disease; while probe sequences or antibodies may be used for diagnosis of such diseases.

***Calcium binding:***

This category contains calcium binding proteins, ligand binding or carriers,  
25       such as diacylglycerol kinase, Calpain, calcium-dependent protein serine/threonine phosphatase, calcium sensing proteins, calcium storage proteins.

***Oxidoreductase:***

This category contains enzymes that catalyze oxidation-reduction reactions, such as oxidoreductases acting on the following groups of donors:  
30       CH-OH, CH-CH, CH-NH<sub>2</sub>, CH-NH; oxidoreductases acting on NADH or

NADPH, nitrogenous compounds, sulfur group of donors, heme group, hydrogen group, diphenols and related substances as donors; oxidoreductases acting on peroxide as acceptor, superoxide radicals as acceptor, oxidizing metal ions, CH<sub>2</sub> groups; oxidoreductases acting on reduced ferredoxin as donor; 5 oxidoreductases acting on reduced flavodoxin as donor; and oxidoreductases acting on the aldehyde or oxo group of donors.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases caused by 10 non-normal activity of oxidoreductases; while probe sequences or antibodies may be used for diagnosis of such diseases.

***Receptors:***

This category contains various receptors, such as signal transducers, complement receptors, ligand-dependent nuclear receptors, transmembrane 15 receptors, GPI-anchored membrane-bound receptors, various coreceptors, internalization receptors, receptors to neurotransmitters, hormones and various other effectors and ligands.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of 20 altering expression of such proteins may be used to treat diseases caused by non-normal activity of oxidoreductases diseases involving various receptors, including receptors to neurotransmitters, hormones and various other effectors and ligands; while probe sequences or antibodies may be used for diagnosis of such diseases.

Examples of these diseases include, but are not limited to, chronic 25 myelomonocytic leukemia caused by growth factor beta receptor deficiency [Rao DS, Chang JC, Kumar PD, Mizukami I, Smithson GM, Bradley SV, Parlow AF, Ross TS (2001) Mol Cell Biol, 21(22):7796-806], thrombosis associated with protease-activated receptor deficiency [Sambrano GR, Weiss EJ, Zheng YW, Huang W, Coughlin SR (2001) Nature, 413(6851):26-7], hypercholesterolemia 30 associated with low density lipoprotein receptor deficiency [Koivisto UM,

Hubbard AL, Mellman I (2001) Cell, 105(5):575-85], familial Hibernian fever associated with tumour necrosis factor receptor deficiency [Simon A, Drenth JP, van der Meer JW (2001) Ned Tijdschr Geneesk, 145(2):77-8], colitis associated with immunoglobulin E receptor expression [Dombrowicz D, Nutten S, Desreumaux P, Neut C, Torpier G, Peeters M, Colombel JF, Capron M (2001) J Exp Med, 193(1):25-34], and alagille syndrome associated with Jagged1 [Stankiewicz P, Rujner J, Loffler C, Kruger A, Nimmakayalu M, Pilacik B, Krajewska-Walasek M, Gutkowska A, Hansmann I, Giannakudis I (2001) Am J Med Genet, 103(2):166-71].

10       ***Protein serine/threonine kinases:***

      This category contains kinases which phosphorylate serine/threonine residues, mainly involved in signal transduction, such as transmembrane receptor protein serine/threonine kinase, 3-phosphoinositide-dependent protein kinase, DNA-dependent protein kinase, G-protein-coupled receptor phosphorylating protein kinase, SNF1A/AMP-activated protein kinase, casein kinase, calmodulin regulated protein kinase, cyclic-nucleotide dependent protein kinase, cyclin-dependent protein kinase, eukaryotic translation initiation factor 2alpha kinase, galactosyltransferase-associated kinase, glycogen synthase kinase 3, protein kinase C, receptor signaling protein serine/threonine kinase, ribosomal protein S6 kinase, Ikb kinase.

      Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases which may be ameliorated by a modulating kinase activity, which is one of the main signaling pathways inside cell; while probe sequences or antibodies may be used for diagnosis of such diseases.

***Channel/pore class transporters:***

      This category contains proteins that mediate the transport of molecules and macromolecules across membranes, such as alpha-type channels, porins, pore-forming toxins.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases in which the transport of molecules and macromolecules such as neurotransmitters, hormones, sugar etc. is non-normal leading to various pathologies; while probe sequences or antibodies may be used for diagnosis of such diseases.

***Hydrolases, acting on acid anhydrides:***

This category contains hydrolytic enzymes that are acting on acid anhydrides, such as hydrolases acting on acid anhydrides, in phosphorus-containing anhydrides, in sulfonyl-containing anhydrides; and hydrolases catalysing transmembrane movement of substances, and involved in cellular and subcellular movement.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases in which the hydrolase-related activities are non-normal (increased or decreased); while probe sequences or antibodies may be used for diagnosis of such diseases.

***Transferases, transferring phosphorus-containing groups:***

This category contains various enzymes that catalyze the transfer of phosphate from one molecule to another, such as phosphotransferases using the following groups as acceptors: alcohol group, carboxyl group, nitrogenous group, phosphate; phosphotransferases with regeneration of donors catalysing intramolecular transfers; diphosphotransferases; nucleotidyltransferase; and phosphotransferases for other substituted phosphate groups.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases in which the transfer of functional group to a modulated moiety is not normal so that a beneficial effect may be achieved by modulation of such transfer; while probe sequences or antibodies may be used for diagnosis of such diseases.

***Phosphoric monoester hydrolases:***

This category contains hydrolytic enzymes that are acting on ester bonds, such as: nuclease, sulfuric ester hydrolase, carboxylic ester hydrolase, thiolester hydrolase, phosphoric monoester hydrolase, phosphoric diester hydrolase, 5 triphosphoric monoester hydrolase, diphosphoric monoester hydrolase, and phosphoric triester hydrolase.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases in which the 10 hydrolytic cleavage of a covalent bond with accompanying addition of water, -H being added to one product of the cleavage and -OH to the other, is not normal so that a beneficial effect may be achieved by modulation of such reaction; while probe sequences or antibodies may be used for diagnosis of such diseases.

***Enzyme inhibitors:***

This category contains inhibitors and suppressors of other proteins and enzymes, such as inhibitors of: kinases, phosphatases, chaperones, guanylate cyclase, DNA gyrase, ribonuclease, proteasome inhibitors, diazepam-binding inhibitor, ornithine decarboxylase inhibitor, GTPase inhibitors, dUTP 20 pyrophosphatase inhibitor, phospholipase inhibitor, proteinase inhibitor, protein biosynthesis inhibitors, alpha-amylase inhibitors.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases in which 25 beneficial effect may be achieved by modulating the activity of inhibitors and suppressors of proteins and enzymes; while probe sequences or antibodies may be used for diagnosis of such diseases.

***Electron transporters:***

This category contains ligand binding or carrier proteins involved in 30 electron transport, such as: flavin-containing electron transporter, cytochromes,



electron donors, electron acceptors, electron carriers, and cytochrome-c oxidases.

***Transferases, transferring glycosyl groups:***

This category contains various enzymes that catalyze the transfer of a chemical group, such as a glycosyl, from one molecule to another. It covers  
5 enzymes such as murein lytic endotransglycosylase E, and sialyltransferase.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases in which  
10 the transfer of a glycosyl chemical group from one molecule to another is not normal so that a beneficial effect may be achieved by modulation of such reaction; while probe sequences or antibodies may be used for diagnosis of such diseases.

***Ligases, forming carbon-oxygen bonds:***

This category contains enzymes that catalyze the linkage between carbon  
15 and oxygen, such as ligase forming aminoacyl-tRNA and related compounds.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases in which the  
20 linkage between carbon and oxygen in an energy dependent process is not normal so that a beneficial effect may be achieved by modulation of such reaction; while probe sequences or antibodies may be used for diagnosis of such diseases.

***Ligases:***

This category contains enzymes that catalyze the linkage of two  
25 molecules, generally utilizing ATP as the energy donor, also called synthetase. It covers enzymes such as beta-alanyl-dopamine hydrolase, carbon-oxygen bonds forming ligase, carbon-sulfur bonds forming ligase, carbon-nitrogen bonds forming ligase, carbon-carbon bonds forming ligase, and phosphoric  
30 ester bonds forming ligase.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins, may be used to treat diseases in which the joining together of two molecules in an energy dependent process is not normal so that a beneficial effect may be achieved by modulation of such reaction; while probe sequences or antibodies may be used for diagnosis of such diseases.

***Hydrolases, acting on glycosyl bonds:***

This category contains hydrolytic enzymes that are acting on glycosyl bonds, such as hydrolases hydrolyzing N-glycosyl compounds, S-glycosyl compounds, and O-glycosyl compounds.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases in which the hydrolase-related activities are non-normal (increased or decreased); while probe sequences or antibodies may be used for diagnosis of such diseases.

***Kinases:***

This category contains kinases, which phosphorylate serine/threonine or tyrosine residues, mainly involved in signal transduction. It covers enzymes such as 2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase, NAD(+) kinase, acetylglutamate kinase, adenosine kinase, adenylate kinase, adenylylsulfate kinase, arginine kinase, aspartate kinase, choline kinase, creatine kinase, cytidylate kinase, deoxyadenosine kinase, deoxycytidine kinase, deoxyguanosine kinase, dephospho-CoA kinase, diacylglycerol kinase, dolichol kinase, ethanolamine kinase, galactokinase, glucokinase, glutamate 5-kinase, glycerol kinase, glycerone kinase, guanylate kinase, hexokinase, homoserine kinase, hydroxyethylthiazole kinase, inositol/phosphatidylinositol kinase, ketohexokinase, mevalonate kinase, nucleoside-diphosphate kinase, pantothenate kinase, phosphoenolpyruvate carboxykinase, phosphoglycerate kinase, phosphomevalonate kinase, protein kinase, pyruvate dehydrogenase (lipoamide) kinase, pyruvate kinase,

ribokinase, ribose-phosphate pyrophosphokinase, selenide,water dikinase, shikimate kinase, thiamine pyrophosphokinase, thymidine kinase, thymidylate kinase, uridine kinase, xylulokinase, 1D-myo-inositol-trisphosphate 3-kinase, phosphofructokinase, pyridoxal kinase, sphinganine kinase, riboflavin kinase, 5 2-dehydro-3-deoxygalactonokinase, 2-dehydro-3-deoxygluconokinase, 4-diphosphocytidyl-2C-methyl-D-erythritol kinase, GTP pyrophosphokinase, L-fuculokinase, L-ribulokinase, L-xylulokinase, isocitrate dehydrogenase (NADP+)] kinase, acetate kinase, allose kinase, carbamate kinase, cobinamide kinase, diphosphate-purine nucleoside kinase, fructokinase, glycerate kinase, 10 hydroxymethylpyrimidine kinase, hygromycin-B kinase, inosine kinase, kanamycin kinase, phosphomethylpyrimidine kinase, phosphoribulokinase, polyphosphate kinase, propionate kinase, pyruvate,water dikinase, rhamnulokinase, tagatose-6-phosphate kinase, tetraacyldisaccharide 4'-kinase, thiamine-phosphate kinase, undecaprenol kinase, uridylate kinase, N- 15 acylmannosamine kinase, D-erythro-sphingosine kinase.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases which may be ameliorated by a modulating kinase activity, which is one of the main signaling 20 pathways inside cell; while probe sequences or antibodies may be used for diagnosis of such diseases.

Examples of these diseases include, but are not limited to, acute lymphoblastic leukemia associated with spleen tyrosine kinase deficiency [Goodman PA, Wood CM, Vassilev A, Mao C, Uckun FM (2001) *Oncogene*, 25 20(30):3969-78), ataxia telangiectasia associated with ATM kinase deficiency (Boulton J (2001) *J Clin Pathol*, 54(7):512-6], congenital haemolytic anaemia associated with erythrocyte pyruvate kinase deficiency [Zanella A, Bianchi P, Fermo E, Iurlo A, Zappa M, Vercellati C, Boschetti C, Baronciani L, Cotton F (2001) *Br J Haematol*, 113(1):43-8], mevalonic aciduria caused by 30 mevalonate kinase deficiency [Houten SM, Koster J, Romeijn GJ, Frenkel J, Di

Rocco M, Caruso U, Landrieu P, Kelley RI, Kuis W, Poll-The BT, Gibson KM, Wanders RJ, Waterham HR (2001) Eur J Hum Genet, 9(4):253-9], and acute myelogenous leukemia associated with over-expressed death-associated protein kinase [Guzman ML, Upchurch D, Grimes B, Howard DS, Rizzieri DA, Luger SM, Phillips GL, Jordan CT (2001) Blood, 97(7):2177-9].

***Nucleotide binding:***

This category contains ligand binding or carrier proteins, involved in physical interaction with a nucleotide - any compound consisting of a nucleoside that is esterified with [ortho]phosphate or an oligophosphate at any hydroxyl group on the glycoside moiety, such as purine nucleotide binding proteins.

***Tubulin binding:***

This category contains binding proteins that bind tubulin, such as microtubule binding proteins.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases which are associated with non-normal tubulin activity or structure. Binding of the products of the genes of this family, or antibodies reactive therewith, can modulate a plurality of tubulin activities as well as change microtubulin structure; while probe sequences or antibodies may be used for diagnosis of such diseases.

Examples of these diseases include, but are not limited to, Alzheimer's disease associated with t-complex polypeptide 1 deficiency [Schuller E, Gulesserian T, Seidl R, Cairns N, Lube G (2001) Life Sci, 69(3):263-70], neurodegeneration associated with apoE deficiency [Masliah E, Mallory M, Ge N, Alford M, Veinbergs I, Roses AD (1995) Exp Neurol, 136(2):107-22], progressive axonopathy associated with disfunctional neurofilaments [Griffiths IR, Kyriakides E, Barrie J (1989) Neuropathol Appl Neurobiol, 15(1):63-74], familial frontotemporal dementia associated with tau deficiency [astor P, Pastor E, Carnero C, Vela R, Garcia T, Amer G, Tolosa E, Oliva R (2001) Ann Neurol, 49(2):263-7],

and colon cancer suppressed by APC [White RL (1997) Pathol Biol (Paris), 45(3):240-4].

***Receptor signaling proteins:***

This category contains receptor proteins involved in signal transduction, such as receptor signaling protein serine/threonine kinase, receptor signaling protein tyrosine kinase, receptor signaling protein tyrosine phosphatase, aryl hydrocarbon receptor nuclear translocator, hematopoietin/interferon-class (D200-domain) cytokine receptor signal transducer, transmembrane receptor protein tyrosine kinase signaling protein, transmembrane receptor protein serine/threonine kinase signaling protein, receptor signaling protein serine/threonine kinase signaling protein, receptor signaling protein serine/threonine phosphatase signaling protein, small GTPase regulatory/interacting protein, receptor signaling protein tyrosine kinase signaling protein, receptor signaling protein serine/threonine phosphatase.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases in which the signal-transduction is non-normal, either as a cause, or as a result of the disease; while probe sequences or antibodies may be used for diagnosis of such diseases.

Examples of these diseases include, but are not limited to, complete hypogonadotropic hypogonadism associated with GnRH receptor deficiency [Kottler ML, Chauvin S, Lahlou N, Harris CE, Johnston CJ, Lagarde JP, Bouchard P, Farid NR, Counis R (2000) J Clin Endocrinol Metab, 85(9):3002-8], severe combined immunodeficiency disease associated with IL-7 receptor deficiency (Puel A, Leonard WJ (2000) Curr Opin Immunol, 12(4):468-73), schizophrenia associated N-methyl-D-aspartate receptor deficiency (Mohn AR, Gainetdinov RR, Caron MG, Koller BH (1999) Cell, 98(4):427-36), Yersinia-associated arthritis associated with tumor necrosis factor receptor p55 deficiency [Zhao YX, Zhang H, Chiu B, Payne U, Inman RD (1999) Arthritis Rheum, 42(8):1662-72], and Dwarfism of Sindh caused by growth hormone-releasing hormone receptor

deficiency [aheshwari HG, Silverman BL, Dupuis J, Baumann G (1998) J Clin Endocrinol Metab, 83(11):4065-74].

***Molecular function unknown:***

This category contains various proteins with unknown molecular  
5 function, such as cell surface antigens.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases in which regulation of the recognition, or participation or bind of cell surface antigens to  
10 other moieties may improve the disease. These diseases include autoimmune diseases, various infectious diseases, cancer diseases which involve non cell surface antigens recognition and activity, etc; while probe sequences or antibodies may be used for diagnosis of such diseases.

***Enzyme activators:***

15 This category contains enzyme regulators, such as activators of: kinases, phosphatases, sphingolipids, chaperones, guanylate cyclase, tryptophan hydroxylase, proteases, phospholipases, caspases, proprotein convertase 2 activator, cyclin-dependent protein kinase 5 activator, superoxide-generating NADPH oxidase activator, sphingomyelin phosphodiesterase activator,  
20 monophenol monooxygenase activator, proteasome activator, GTPase activator.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases in which beneficial effect may be achieved by modulating the activity of activators of  
25 proteins and enzymes; while probe sequences or antibodies may be used for diagnosis of such diseases.

***Transferases, transferring one-carbon groups:***

This category contains various enzymes that catalyze the transfer of a chemical group, such as a one-carbon, from one molecule to another. The  
30 category covers enzymes such as methyltransferase, amidinotransferase,

hydroxymethyl-, formyl- and related transferase, carboxyl- and carbamoyltransferase.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases in which the transfer of a one-carbon chemical group from one molecule to another is not normal so that a beneficial effect may be achieved by modulation of such reaction; while probe sequences or antibodies may be used for diagnosis of such diseases.

10       ***Transferases:***

This category contains various enzymes that catalyze the transfer of a chemical group, such as a phosphate or amine, from one molecule to another. It covers enzymes such as: transferases, transferring one-carbon groups, aldehyde or ketonic groups, acyl groups, glycosyl groups, alkyl or aryl (other than methyl) groups, nitrogenous, phosphorus-containing groups, sulfur-containing groups, lipoyltransferase, deoxycytidyl transferases.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases in which the transfer of a chemical group from one molecule to another is not normal so that a beneficial effect may be achieved by modulation of such reaction; while probe sequences or antibodies may be used for diagnosis of such diseases.

***Chaperone:***

This category contains functional classes of unrelated families of proteins that assist the correct non-covalent assembly of other polypeptide-containing structures in vivo, but are not components of these assembled structures when they are performing their normal biological function. The category covers proteins such as: ribosomal chaperone, peptidylprolyl isomerase, lectin-binding chaperone, nucleosome assembly chaperone, chaperonin ATPase, cochaperone, heat shock protein, HSP70/HSP90

organizing protein, fimbrial chaperone, metallochaperone, tubulin folding, HSC70-interacting protein.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of  
5 altering expression of such proteins may be used to treat diseases which are associated with non-normal protein activity or structure or abnormal degradation of such proteins; while probe sequences or antibodies may be used for diagnosis of such diseases.

***Cell adhesion molecule:***

10 This category contains proteins that serve as adhesion molecules between adjoining cells, such as: membrane-associated protein with guanylate kinase activity, cell adhesion receptor, neuroligin, calcium-dependent cell adhesion molecule, selectin, calcium-independent cell adhesion molecule, extracellular matrix protein.

15 Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases in which adhesion between adjoining cells is involved, typically conditions in which the adhesion is non-normal; while probe sequences or antibodies may be used for  
20 diagnosis of such diseases. Typical examples of such conditions are cancer conditions in which non-normal adhesion may cause and enhance the process of metastasis. Other examples of such conditions include conditions of non-normal growth and development of various tissues in which modulation adhesion among adjoining cells can improve the condition.

25 Examples of these diseases include, but are not limited to, Wiskott-Aldrich syndrome associated with WAS deficiency [Westerberg L, Greicius G, Snapper SB, Aspenstrom P, Severinson E (2001) Blood, 98(4):1086-94], asthma associated with intercellular adhesion molecule-1 deficiency [Tang ML, Fiscus LC (2001) Pulm Pharmacol Ther, 14(3):203-10], intra-atrial thrombogenesis associated with  
30 increased von Willebrand factor activity [Fukuchi M, Watanabe J, Kumagai K,



Katori Y, Baba S, Fukuda K, Yagi T, Iguchi A, Yokoyama H, Miura M, Kagaya Y, Sato S, Tabayashi K, Shirato K (2001) J Am Coll Cardiol, 37(5):1436-42], junctional epidermolysis bullosa associated with laminin 5 beta3 deficiency [Robbins PB, Lin Q, Goodnough JB, Tian H, Chen X, Khavari PA (2001) Proc Natl Acad Sci U S A, 98(9):5193-8], and hydrocephalus caused by neural adhesion molecule L1 deficiency [Rolf B, Kutsche M, Bartsch U (2001) Brain Res, 891(1-2):247-52].

***Motor proteins:***

This category contains proteins that are held to generate force or energy by the hydrolysis of ATP and that functions in the production of intracellular movement or transportation. It covers proteins such as: microfilament motor, axonemal motor, microtubule motor, and kinetochore motor (like dynein, kinesin, or myosin).

***Defense/immunity proteins:***

This category contains proteins that are involved in the immune and complement systems, such as acute-phase response proteins, antimicrobial peptides, antiviral response proteins, blood coagulation factors, complement components, immunoglobulins, major histocompatibility complex antigens, opsonins.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases involving the immunological system including inflammation, autoimmune diseases, infectious diseases, as well as cancerous processes or diseases which are manifested by non-normal coagulation processes, which may include abnormal bleeding or excessive coagulation.; while probe sequences or antibodies may be used for diagnosis of such diseases.

Examples of these diseases include, but are not limited to, late (C5-9) complement component deficiency associated with opsonin receptor allotypes [Fijen CA, Bredius RG, Kuijper EJ, Out TA, De Haas M, De Wit AP, Daha MR,

De Winkel JG (2000) Clin Exp Immunol, 120(2):338-45], combined immunodeficiency associated with defective expression of MHC class II genes [Griscelli C, Lisowska-Grospierre B, Mach B (1989) Immunodef Rev 1(2):135-53], loss of antiviral activity of CD4 T cells caused by neutralization of endogenous TNF alpha [Pavic I, Polic B, Crnkovic I, Lucin P, Jonjic S, Koszinowski UH (1993) J Gen Virol, 74 (Pt 10):2215-23], autoimmune diseases associated with natural resistance-associated macrophage protein deficiency [Evans CA, Harbuz MS, Ostensfeld T, Norrish A, Blackwell JM (2001) Neurogenetics, 3(2):69-78], and Epstein-Barr virus-associated lymphoproliferative disease inhibited by combined GM-CSF and IL-2 therapy [Baiocchi RA, Ward JS, Carrodegua L, Eisenbeis CF, Peng R, Roychowdhury S, Vourganti S, Sekula T, O'Brien M, Moeschberger M, Caligiuri MA (2001) J Clin Invest, 108(6):887-94].

#### ***Intracellular transporters:***

This category contains proteins that mediate the transport of molecules and macromolecules inside the cell, such as: intracellular nucleoside transporter, vacuolar assembly proteins, vesicle transporters, vesicle fusion proteins, type II protein secretors.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases in which the transport of molecules and macromolecules is non-normal leading to various pathologies; while probe sequences or antibodies may be used for diagnosis of such diseases.

#### ***Transporters:***

This category contains proteins that mediate the transport of molecules and macromolecules, such as channels, exchangers, pumps. The category covers proteins such as: amine/polyamine transporter, lipid transporter, neurotransmitter transporter, organic acid transporter, oxygen transporter, water transporter, carriers, intracellular transporters, protein transporters, ion transporters, carbohydrate transporter, polyol transporter, amino acid

transporters, vitamin/cofactor transporters, siderophore transporter, drug transporter, channel/pore class transporter, group translocator, auxiliary transport proteins, permeases, murein transporter, organic alcohol transporter, nucleobase, nucleoside, nucleotide and nucleic acid transporters.

5        Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases in which the transport of molecules and macromolecules such as neurotransmitters, hormones, sugar etc. is non-normal leading to various pathologies; while probe sequences or  
10        antibodies may be used for diagnosis of such diseases.

      Examples of these diseases include, but are not limited to, glycogen storage disease caused by glucose-6-phosphate transporter deficiency (Hiraiwa H, Chou JY (2001) DNA Cell Biol, 20(8):447-53), tangier disease associated with ATP-binding cassette transporter-1 deficiency (McNeish J, Aiello RJ, Guyot D, Turi T, Gabel C, Aldinger C, Hoppe KL, Roach ML, Royer LJ, de Wet J, Broccardo C, Chimini G, Francone OL (2000) Proc Natl Acad Sci U S A, 97(8):4245-50),  
15        systemic primary carnitine deficiency associated with organic cation transporter deficiency (Tang NL, Ganapathy V, Wu X, Hui J, Seth P, Yuen PM, Wanders RJ, Fok TF, Hjelm NM (1999) Hum Mol Genet, 8(4):655-60), Wilson disease  
20        associated with copper-transporting ATPases deficiency (Payne AS, Kelly EJ, Gitlin JD (1998) Proc Natl Acad Sci U S A, 95(18):10854-9), and atelosteogenesis associated with diastrophic dysplasia sulphate transporter deficiency (Newbury-Ecob R (1998) J Med Genet, 35(1):49-53).

#### ***Lyases:***

25        This category contains enzymes that catalyze the formation of double bonds by removing chemical groups from a substrate without hydrolysis or catalyze the addition of chemical groups to double bonds. It covers enzymes such as carbon-carbon lyase, carbon-oxygen lyase, carbon-nitrogen lyase, carbon-sulfur lyase, carbon-halide lyase, phosphorus-oxygen lyase, and other  
30        lyases.

***Actin binding proteins:***

This category contains actin binding proteins, such as actin cross-linking, actin bundling, F-actin capping, actin monomer binding, actin lateral binding, actin depolymerizing, actin monomer sequestering, actin filament severing, 5 actin modulating, membrane associated actin binding, actin thin filament length regulation, and actin polymerizing proteins.

***Protein binding proteins:***

This category contains various proteins, involved in diverse biological functions, such as: intermediate filament binding, LIM-domain binding, LLR- 10 domain binding, clathrin binding, ARF binding, vinculin binding, KU70 binding, troponin C binding PDZ-domain binding, SH3-domain binding, fibroblast growth factor binding, membrane-associated protein with guanylate kinase activity interacting, Wnt-protein binding, DEAD/H-box RNA helicase binding, beta-amyloid binding, myosin binding, TATA-binding protein binding 15 DNA topoisomerase I binding, polypeptide hormone binding, RHO binding, FH1-domain binding, syntaxin-1 binding, HSC70-interacting, transcription factor binding, metarhodopsin binding, tubulin binding, JUN kinase binding, RAN protein binding, protein signal sequence binding, importin alpha export receptor, poly-glutamine tract binding, protein carrier, beta-catenin binding, 20 protein C-terminus binding, lipoprotein binding, cytoskeletal protein binding protein, nuclear localization sequence binding, protein phosphatase 1 binding, adenylate cyclase binding, eukaryotic initiation factor 4E binding, calmodulin binding, collagen binding, insulin-like growth factor binding, lamin binding, profilin binding, tropomyosin binding, actin binding, peroxisome targeting 25 sequence binding, SNARE binding, cyclin binding.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases which are associated with non-normal protein activity or structure. Binding of the products of 30 the variants of this family, or antibodies reactive therewith, can modulate a

plurality of protein activities as well as change protein structure; while probe sequences or antibodies may be used for diagnosis of such diseases.

***Ligand binding or carrier proteins:***

This category contains various proteins, involved in diverse biological functions, such as: pyridoxal phosphate binding, carbohydrate binding, magnesium binding, amino acid binding, cyclosporin A binding, nickel binding, chlorophyll binding, biotin binding, penicillin binding, selenium binding, tocopherol binding, lipid binding, drug binding, oxygen transporter, electron transporter, steroid binding, juvenile hormone binding, retinoid binding, heavy metal binding, calcium binding, protein binding, glycosaminoglycan binding, folate binding, odorant binding, lipopolysaccharide binding, nucleotide binding.

***ATPases:***

This category contains enzymes that catalyze the hydrolysis of ATP to ADP, releasing energy that is used in the cell; adenosine triphosphatase. It covers enzymes such as plasma membrane cation-transporting ATPase, ATP-binding cassette (ABC) transporter, magnesium-ATPase, hydrogen-/sodium-translocating ATPase, arsenite-transporting ATPase, protein-transporting ATPase, DNA translocase, P-type ATPase, hydrolase, acting on acid anhydrides, - involved in cellular and subcellular movement.

***Carboxylic ester hydrolases:***

This category contains hydrolytic enzymes, acting on carboxylic ester bonds, such as N-acetylglucosaminylphosphatidylinositol deacetylase, 2-acetyl-1-alkylglycerophosphocholine esterase, aminoacyl-tRNA hydrolase, arylesterase, carboxylesterase, cholinesterase, gluconolactonase, sterol esterase, acetylesterase, carboxymethylenebutenolidase, protein-glutamate methylesterase, lipase, 6-phosphogluconolactonase.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases in which the hydrolytic cleavage of a covalent bond with accompanying addition of water, -H

being added to one product of the cleavage and -OH to the other, is not normal so that a beneficial effect may be achieved by modulation of such reaction; while probe sequences or antibodies may be used for diagnosis of such diseases.

***Hydrolase, acting on ester bonds:***

5        This category contains hydrolytic enzymes, acting on ester bonds, such as nucleases, sulfuric ester hydrolase, carboxylic ester hydrolases, thiolester hydrolase, phosphoric monoester hydrolase, phosphoric diester hydrolase, triphosphoric monoester hydrolase, diphosphoric monoester hydrolase, phosphoric triester hydrolase.

10        Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases in which the hydrolytic cleavage of a covalent bond with accompanying addition of water, -H being added to one product of the cleavage and -OH to the other, is not normal so  
15        that a beneficial effect may be achieved by modulation of such reaction; while probe sequences or antibodies may be used for diagnosis of such diseases.

***Hydrolases:***

      This category contains hydrolytic enzymes, such as GPI-anchor transamidase, peptidases, hydrolases, acting on ester bonds, glycosyl bonds,  
20        ether bonds, carbon-nitrogen (but not peptide) bonds, acid anhydrides, acid carbon-carbon bonds, acid halide bonds, acid phosphorus-nitrogen bonds, acid sulfur-nitrogen bonds, acid carbon-phosphorus bonds, acid sulfur-sulfur bonds.

      Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of  
25        altering expression of such proteins may be used to treat diseases in which the hydrolytic cleavage of a covalent bond with accompanying addition of water, -H being added to one product of the cleavage and -OH to the other, is not normal so that a beneficial effect may be achieved by modulation of such reaction; while probe sequences or antibodies may be used for diagnosis of such diseases.

***Enzymes:***

This category contains naturally occurring or synthetic macromolecular substance composed wholly or largely of protein, that catalyzes, more or less specifically, one or more (bio)chemical reactions at relatively low temperatures.

- 5 The action of RNA that has catalytic activity (ribozyme) is often also regarded as enzymic. Nevertheless, enzymes are mainly proteinaceous and are often easily inactivated by heating or by protein-denaturing agents. The substances upon which they act are known as substrates, for which the enzyme possesses a specific binding or active site.

- 10 This category covers various proteins possessing enzymatic activities, such as mannosylphosphate transferase, para-hydroxybenzoate:polyprenyltransferase, Rieske iron-sulfur protein, imidazoleglycerol-phosphate synthase, sphingosine hydroxylase, tRNA 2'-phosphotransferase, sterol C-24(28) reductase, C-8 sterol isomerase, C-22 sterol  
15 desaturase, C-14 sterol reductase, C-3 sterol dehydrogenase (C-4 sterol decarboxylase), 3-keto sterol reductase, C-4 methyl sterol oxidase, dihydronicotinamide riboside quinone reductase, glutamate phosphate reductase, DNA repair enzyme, telomerase, alpha-ketoacid dehydrogenase, beta-alanyl-dopamine synthase, RNA editase, aldo-keto reductase, alkylbase  
20 DNA glycosidase, glycogen debranching enzyme, dihydropterin deaminase, dihydropterin oxidase, dimethylnitrosamine demethylase, ecdysteroid UDP-glucosyl/UDP glucuronosyl transferase, glycine cleavage system, helicase, histone deacetylase, mevaldate reductase, monooxygenase, poly(ADP-ribose) glycohydrolase, pyruvate dehydrogenase, serine esterase, sterol carrier protein  
25 X-related thiolase, transposase, tyramine-beta hydroxylase, para-aminobenzoic acid (PABA) synthase, glu-tRNA(gln) amidotransferase, molybdopterin cofactor sulfurase, lanosterol 14-alpha-demethylase, aromatase, 4-hydroxybenzoate octaprenyltransferase, 7,8-dihydro-8-oxoguanine-triphosphatase, CDP-alcohol phosphotransferase, 2,5-diamino-6-  
30 (ribosylamino)-4(3H)-pyrimidonone 5'-phosphate deaminase, diphosphoinositol

polyphosphate phosphohydrolase, gamma-glutamyl carboxylase, small protein conjugating enzyme, small protein activating enzyme, 1-deoxyxylulose-5-phosphate synthase, 2'-phosphotransferase, 2-octoprenyl-3-methyl-6-methoxy-1,4-benzoquinone hydroxylase, 2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase, 3,4 dihydroxy-2-butanone-4-phosphate synthase, 4-amino-4-deoxychorismate lyase, 4-diphosphocytidyl-2C-methyl-D-erythritol synthase, ADP-L-glycero-D-manno-heptose synthase, D-erythro-7,8-dihydroneopterin triphosphate 2'-epimerase, N-ethylmaleimide reductase, O-antigen ligase, O-antigen polymerase, UDP-2,3-diacylglucosamine hydrolase, arsenate reductase, carnitine racemase, cobalamin [5'-phosphate] synthase, cobinamide phosphate guanylyltransferase, enterobactin synthetase, enterochelin esterase, enterochelin synthetase, glycolate oxidase, integrase, lauroyl transferase, peptidoglycan synthetase, phosphopantetheinyltransferase, phosphoglucosamine mutase, phosphoheptose isomerase, quinolinate synthase, siroheme synthase, N-acylmannosamine-6-phosphate 2-epimerase, N-acetyl-anhydromuramoyl-L-alanine amidase, carbon-phosphorous lyase, heme-copper terminal oxidase, disulfide oxidoreductase, phthalate dioxygenase reductase, sphingosine-1-phosphate lyase, molybdopterin oxidoreductase, dehydrogenase, NADPH oxidase, naringenin-chalcone synthase, N-ethylammelene chlorohydrolase, polyketide synthase, aldolase, kinase, phosphatase, CoA-ligase, oxidoreductase, transferase, hydrolase, lyase, isomerase, ligase, ATPase, sulfhydryl oxidase, lipoate-protein ligase, delta-1-pyrroline-5-carboxyate synthetase, lipoic acid synthase, and tRNA dihydrouridine synthase.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases which can be ameliorated by modulating the activity of various enzymes which are involved both in enzymatic processes inside cells as well as in cell signaling; while probe sequences or antibodies may be used for diagnosis of such diseases.



***Cytoskeletal proteins:***

This category contains proteins involved in the structure formation of the cytoskeleton.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases which are caused or due to abnormalities in cytoskeleton, including cancerous cells, and diseased cells including those which do not propagate, grow or function normally; while probe sequences or antibodies may be used for diagnosis of such diseases.

***Structural proteins:***

This category contains proteins involved in the structure formation of the cell, such as: structural proteins of ribosome, cell wall structural proteins, structural proteins of cytoskeleton, extracellular matrix structural proteins, extracellular matrix glycoproteins, amyloid proteins, plasma proteins, structural proteins of eye lens, structural protein of chorion (sensu Insecta), structural protein of cuticle (sensu Insecta), puparial glue protein (sensu Diptera), structural proteins of bone, yolk proteins, structural proteins of muscle, structural protein of vitelline membrane (sensu Insecta), structural proteins of peritrophic membrane (sensu Insecta), structural proteins of nuclear pores.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases which are caused or due to abnormalities in cytoskeleton, including cancerous cells, and diseased cells including those which do not propagate, grow or function normally; while probe sequences or antibodies may be used for diagnosis of such diseases.

***Ligands:***

This category contains proteins that bind to another chemical entity to form a larger complex, involved in various biological processes, such as signal transduction, metabolism, growth and differentiation, etc. The category covers ligands such as: opioid peptides, baboon receptor ligand, branchless receptor

ligand, breathless receptor ligand, ephrin, frizzled receptor ligand, frizzled-2 receptor ligand, heartless receptor ligand, Notch receptor ligand, patched receptor ligand, punt receptor ligand, Ror receptor ligand, saxophone receptor ligand, SE20 receptor ligand, sevenless receptor ligand, smooth receptor ligand, 5 thickveins receptor ligand, Toll receptor ligand, Torso receptor ligand, death receptor ligand, scavenger receptor ligand, neuroligin, integrin ligand, hormones, pheromones, growth factors, sulfonylurea receptor ligand.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of 10 altering expression of such proteins may be used to treat diseases which involve non-normal secretion of proteins which may be due to non-normal presence, absence or non-normal response to normal levels of secreted proteins including hormones, neurotransmitters, and various other proteins secreted by cells to the extracellular environment or diseases which are endocrine in nature (cause or are a 15 result of hormones); while probe sequences or antibodies may be used for diagnosis of such diseases.

Examples of these diseases include, but are not limited to, analgesia inhibited by orphanin FQ/nociceptin [Shane R, Lazar DA, Rossi GC, Pasternak GW, Bodnar RJ (2001) *Brain Res*, 907(1-2):109-16], stroke protected by estrogen 20 [Alkayed NJ, Goto S, Sugo N, Joh HD, Klaus J, Crain BJ, Bernard O, Traystman RJ, Hum PD (2001) *J Neurosci*, 21(19):7543-50], atherosclerosis associated with growth hormone deficiency [Elhadd TA, Abdu TA, Oxtoby J, Kennedy G, McLaren M, Neary R, Belch JJ, Clayton RN (2001) *J Clin Endocrinol Metab*, 86(9):4223-32], diabetes inhibited by alpha-galactosylceramide [Hong S, Wilson 25 MT, Serizawa I, Wu L, Singh N, Naidenko OV, Miura T, Haba T, Scherer DC, Wei J, Kronenberg M, Koezuka Y, Van Kaer L (2001) *Nat Med*, 7(9):1052-6], and Huntington's disease associated with huntingtin deficiency [Rao DS, Chang JC, Kumar PD, Mizukami I, Smithson GM, Bradley SV, Parlow AF, Ross TS (2001) *Mol Cell Biol*, 21(22):7796-806].

***Signal transducer:***

This category contains various signal transducers, such as: activin inhibitors, receptor-associated proteins, alpha-2 macroglobulin receptors, morphogens, quorum sensing signal generators, quorum sensing response  
5 regulators, receptor signaling proteins, ligands, receptors, two-component sensor molecules, two-component response regulators.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases in which the  
10 signal-transduction is non-normal, either as a cause, or as a result of the disease; while probe sequences or antibodies may be used for diagnosis of such diseases.

Examples of these diseases include, but are not limited to, altered sexual dimorphism associated with signal transducer and activator of transcription 5b [Udy GB, Towers RP, Snell RG, Wilkins RJ, Park SH, Ram PA, Waxman DJ,  
15 Davey HW (1997) Proc Natl Acad Sci U S A, 94(14):7239-44], multiple sclerosis associated with sgp130 deficiency [Padberg F, Feneberg W, Schmidt S, Schwarz MJ, Korschenhausen D, Greenberg BD, Nolde T, Muller N, Trapmann H, Konig N, Moller HJ, Hampel H (1999) J Neuroimmunol, 99(2):218-23], intestinal inflammation associated with elevated signal transducer and activator of  
20 transcription 3 activity [Suzuki A, Hanada T, Mitsuyama K, Yoshida T, Kamizono S, Hoshino T, Kubo M, Yamashita A, Okabe M, Takeda K, Akira S, Matsumoto S, Toyonaga A, Sata M, Yoshimura A (2001) J Exp Med, 193(4):471-81], carcinoid tumor inhibited by increased signal transducer and activators of transcription 1 and  
2 [Zhou Y, Wang S, Gobl A, Oberg K (2001) Oncology, 60(4):330-8], and  
25 esophageal cancer associated with loss of EGF-STAT1 pathway [Watanabe G, Kaganoi J, Imamura M, Shimada Y, Itami A, Uchida S, Sato F, Kitagawa M (2001) Cancer J, 7(2):132-9].

***RNA polymerase II transcription factors:***

This category contains proteins, such as specific and non-specific RNA  
30 polymerase II transcription factors, enhancer binding, ligand-regulated

transcription factor, general RNA polymerase II transcription factors.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases involving RNA polymerase II transcription factors, for example diseases where there is non-normal transcription of RNA; while probe sequences or antibodies may be used for diagnosis of such diseases.

***RNA binding proteins:***

This category contains RNA binding proteins involved in splicing and translation regulation, such as tRNA binding proteins, RNA helicases, double-stranded RNA and single-stranded RNA binding proteins, mRNA binding proteins, snRNA cap binding proteins, 5S RNA and 7S RNA binding proteins, poly-pyrimidine tract binding proteins, snRNA binding proteins, and AU-specific RNA binding proteins.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases involving transcription and translation factors such as: helicases, isomerases, histones and nucleases, for example diseases where there is non-normal transcription, splicing, post-transcriptional processing, translation or stability of the RNA; while probe sequences or antibodies may be used for diagnosis of such diseases.

***Nucleic acid binding proteins:***

This category contains proteins involved in RNA and DNA synthesis and expression regulation, such as transcription factors, RNA and DNA binding proteins, zinc fingers, helicase, isomerase, histones, nucleases, ribonucleoproteins, transcription and translation factors and other.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat diseases involving DNA or RNA binding proteins such as: helicases, isomerases, histones and nucleases,

for example diseases where there is non-normal replication or transcription of DNA and RNA respectively; while probe sequences or antibodies may be used for diagnosis of such diseases.

***Proteins involved in Metabolism:***

5       The totality of the chemical reactions and physical changes that occur in living organisms, comprising anabolism and catabolism; may be qualified to mean the chemical reactions and physical processes undergone by a particular substance, or class of substances, in a living organism.

10       This category covers proteins involved in the reactions of cell growth and maintenance, such as: metabolism resulting in cell growth, carbohydrate metabolism, energy pathways, electron transport, nucleobase, nucleoside, nucleotide and nucleic acid metabolism, protein metabolism and modification, amino acid and derivative metabolism, protein targeting, lipid metabolism, aromatic compound metabolism, one-carbon compound metabolism,  
15       coenzymes and prosthetic group metabolism, sulfur metabolism, phosphorus metabolism, phosphate metabolism, oxygen and radical metabolism, xenobiotic metabolism, nitrogen metabolism, fat body metabolism (sensu Insecta), protein localization, catabolism, biosynthesis, toxin metabolism, methylglyoxal metabolism, cyanate metabolism, glycolate metabolism, carbon utilization,  
20       antibiotic metabolism.

      Examples of metabolism-related diseases include, but are not limited to, multisystem mitochondrial disorder caused by mitochondrial DNA cytochrome C oxidase II deficiency [Campos Y, Garcia-Redondo A, Fernandez-Moreno MA, Martinez-Pardo M, Goda G, Rubio JC, Martin MA, del Hoyo P, Cabello  
25       A, Bornstein B, Garesse R, Arenas J (2001) Ann Neurol Sep;50(3):409-13], conduction defects and ventricular dysfunction in the heart associated with heterogeneous connexin43 expression [Gutstein DE, Morley GE, Vaidya D, Liu F, Chen FL, Stuhlmann H, Fishman GI (2001) Circulation, 104(10):1194-9], atherosclerosis associated with growth suppressor p27 deficiency [Diez-Juan A,  
30       Andres V (2001) FASEB J, 15(11):1989-95], colitis associated with glutathione

peroxidase deficiency [Esworthy RS, Aranda R, Martin MG, Doroshow JH, Binder SW, Chu FF (2001) Am J Physiol Gastrointest Liver Physiol, 281(3):G848-55], and systemic lupus erythematosus associated with deoxyribonuclease I deficiency [Yasutomo K, Horiuchi T, Kagami S, Tsukamoto H, Hashimura C, Urushihara M, Kuroda Y (2001) Nat Genet, 28(4):313-4].

***Cell growth and/or maintenance proteins:***

This category contains proteins involved in any biological process required for cell survival, growth and maintenance. It covers proteins involved in biological processes such as: cell organization and biogenesis, cell growth, cell proliferation, metabolism, cell cycle, budding, cell shape and cell size control, sporulation (sensu Saccharomyces), transport, ion homeostasis, autophagy, cell motility, chemi-mechanical coupling, membrane fusion, cell-cell fusion, stress response.

Pharmaceutical compositions including such proteins or protein encoding sequences, antibodies directed against such proteins or polynucleotides capable of altering expression of such proteins may be used to treat or prevent diseases such as cancer, degenerative diseases, for example neurodegenerative diseases or conditions associated with aging, or alternatively, diseases wherein apoptosis which should have taken place, does not take place; while probe sequences or antibodies may be used for diagnosis of such diseases. Detection of pre-disposition to a disease, as well as for determination of the stage of the disease can also be effected

Examples of these diseases include, but are not limited to, ataxia-telangiectasia associated with ataxia-telangiectasia mutated deficiency [Hande et al (2001) Hum Mol Genet, 10(5):519-28], osteoporosis associated with osteonectin deficiency [Delany et al (2000) J Clin Invest, 105(7):915-23], arthritis caused by membrane-bound matrix metalloproteinase deficiency [Holmbeck et al (1999) Cell, 99(1):81-92], defective stratum corneum and early neonatal death associated with transglutaminase 1 deficiency [Matsuki et al

(1998) Proc Natl Acad Sci U S A, 95(3):1044-9], and Alzheimer's disease associated with estrogen [Simpkins et al (1997) Am J Med, 103(3A):19S-25S].

Thus, the nucleic acid sequences of the present invention and the proteins encoded thereby and the cells and antibodies described hereinabove can be used in, for example, screening assays, therapeutic or prophylactic methods of treatment, or predictive medicine (e.g., diagnostic and prognostic assays, including those used to monitor clinical trials, and pharmacogenetics).

More specifically, the nucleic acids of the invention can be used to: (i) express a protein of the invention in a host cell (in culture or in an intact multicellular organism following, e.g., gene therapy, given, of course, that the transcript in question contains more than untranslated sequence); (ii) detect an mRNA; or (iii) detect an alteration in a gene to which a nucleic acid of the invention specifically binds; or to modulate such a gene's activity.

The nucleic acids and proteins of the invention can also be used to treat disorders characterized by either insufficient or excessive production of those nucleic acids or proteins, a failure in a biochemical pathway in which they normally participate in a cell, or other aberrant or unwanted activity relative to the wild type protein (e.g., inappropriate enzymatic activity or unproductive protein folding). The proteins of the invention are especially useful in screening for naturally occurring protein substrates or other compounds (e.g., drugs) that modulate protein activity. The antibodies of the invention can also be used to detect and isolate the proteins of the invention, to regulate their bioavailability, or otherwise modulate their activity. These uses, and the methods by which they can be achieved, are described in detail below.

### *Screening Assays*

The present invention provides methods (or "screening assays") for identifying agents (or "test compounds" that bind to or otherwise modulate (i.e., stimulate or inhibit) the expression or activity of a nucleic acid of the present invention or the protein it encodes. An agent may be, for example, a small molecule such as a peptide, peptidomimetic (e.g., a peptoid), an amino acid or

an analog thereof, a polynucleotide or an analog thereof, a nucleotide or an analog thereof, or an organic or inorganic compound (e.g., a heteroorganic or organometallic compound) having a molecular weight less than about 10,000 (e.g., about 5,000, 1,000, or 500) grams per mole and salts, esters, and other  
5 pharmaceutically acceptable forms of such compounds.

Agents identified in the screening assays can be used, for example, to modulate the expression or activity of the nucleic acids or proteins of the invention in a therapeutic protocol, or to discover more about the biological functions of the proteins.

10 The assays can be constructed to screen for agents that modulate the expression or activity of a protein of the invention or another cellular component with which it interacts. For example, where the protein of the invention is an enzyme, the screening assay can be constructed to detect agents that modulate either the enzyme's expression or activity or that of its substrate.

15 The agents tested can be those obtained from combinatorial libraries. Methods known in the art allow the production and screening of: biological libraries; peptoid libraries [i.e., libraries of molecules that function as peptides even though they have a non-peptide backbone that confers resistance to enzymatic degradation; see, e.g., Zuckermann et al., J. Med. Chem. 37:2678-85, (1994)];

20 spatially addressable parallel solid phase or solution phase libraries; synthetic libraries requiring deconvolution; "one-bead one-compound" libraries; and synthetic libraries. The biological and peptoid libraries can be used to test only peptides, but the other four are applicable to testing peptides, non-peptide oligomers or libraries of small molecules [Lam, Anticancer Drug Des. 12:145, (1997)].

25 Molecular libraries can be synthesized as described by DeWitt et al. [Proc. Natl. Acad. Sci. USA 90:6909, (1993)] Erb et al. [Proc. Natl. Acad. Sci. USA 91:11422, (1994)] Zuckermann et al. [J. Med. Chem. 37:2678, (1994)] Cho et al. [Science 261:1303, (1993)] and Gallop et al. [J. Med. Chem. 37:1233, (1994)].



Libraries of compounds may be presented in solution [see, e.g., Houghten, *Biotechniques* 13:412-421, (1992)], or on beads [Lam, *Nature* 354:82-84, (1991)], chips [Fodor, *Nature* 364:555-556, (1993)], bacteria or spores (U.S. Patent No. 5,223,409), plasmids [Cull et al., *Proc Natl Acad Sci USA* 89:1865-1869, (1992)] or on phage. [Scott and Smith, *Science* 249:386-390, (1990); Devlin, *Science* 249:404-406, (1990); Cwirla et al., *Proc. Natl. Acad. Sci. USA* 87:6378-6382, (1990); Felici, *J. Mol. Biol.* 222:301-310, (1991); and U.S. Patent No. 5,223,409].

The screening assay can be a cell-based assay, in which case the screening method includes contacting a cell that expresses a protein of the invention with a test compound and determining the ability of the test compound to modulate the protein's activity. The cell used can be a mammalian cell, including a cell obtained from a human or from a human cell line.

Alternatively, or in addition to examining the ability of an agent to modulate expression or activity generally, one can examine the ability of an agent to interact with, for example, to specifically bind to, a nucleic acid or protein of the invention. For example, one can couple an agent (e.g., a substrate) to a label (those described above, including radioactive or enzymatically active substances, are suitable), contact the nucleic acid or protein of the invention with the labeled agent, and determine whether they bind one another (by detecting, for example, a complex containing the nucleic acid or protein and the labeled agent). Labels are not, however, always required. For example, one can use a microphysiometer to detect interaction between an agent and a protein of the invention, neither of which were previously labeled [McConnell et al., *Science* 257:1906-1912, (1992)]. A microphysiometer (also known as a cytosensor) is an analytical instrument that measures the rate at which a cell acidifies its environment. The instrument uses a light-addressable potentiometric sensor (LAPS), and changes in the acidification rate indicate interaction between an agent and a protein of the invention. Molecular

interactions can also be detected using fluorescence energy transfer (FET; see, e.g., U.S. Patent Nos. 5,631,169 and 4,868,103). An FET binding event can be conveniently measured through fluorometric detection means well known in the art (e.g., by means of a fluorimeter). Where analysis in real time is desirable, one can examine the interaction (e.g., binding) between an agent and a protein of the invention with Biomolecular Interaction Analysis [BIA; see, e.g., Sjolander and Urbaniczky *Anal. Chem.* 63:2338-2345, (1991) and Szabo et al., *Curr. Opin. Struct. Biol.* 5:699-705, (1995)]. BIA allows one to detect biospecific interactions in real time without labeling any of the interactants (e.g., BIAcore).

The screening assays can also be cell-free assays (i.e., soluble or membrane-bound forms of the proteins of the invention, including the variants, mutants, and other fragments described above, can be used to identify agents that bind those proteins or otherwise modulate their expression or activity). The basic protocol is the same as that for a cell-based assay in that, in either case, one must contact the protein of the invention with an agent of interest [for a sufficient time and under appropriate (e.g., physiological) conditions] to allow any potential interaction to occur and then determine whether the agent binds the protein or otherwise modulates its expression or activity.

Those of ordinary skill in the art will, however, appreciate that there are differences between cell-based and cell-free assays. For example, when membrane-bound forms of the protein are used, it may be desirable to utilize a solubilizing agent (e.g., non-ionic detergents such as n-octylglucoside, n-dodecylglucoside, n-dodecylmaltoside, octanoyl-N-methylglucamide, decanoyl-N-methylglucamide, Triton<sup>®</sup> X-100, Triton<sup>®</sup> X-114, Thesit<sup>®</sup>, Isotridecypoly(ethylene glycol ether)<sub>n</sub>, 3-[(3-cholamidopropyl)dimethylamminio]-1-propane sulfonate (CHAPS), 3-[(3-cholamidopropyl)dimethylamminio]-2-hydroxy-1-propane sulfonate (CHAPSO), or N-dodecyl=N,N-dimethyl-3-ammonio-1-propane sulfonate).

In the assays of the invention, any of the proteins described herein or the agents being tested can be anchored to a solid phase or otherwise immobilized (assays in which one of two substances that interact with one another are anchored to a solid phase are sometimes referred to as "heterogeneous" assays).

5 For example, a protein of the present invention can be anchored to a microtiter plate, a test tube, a microcentrifuge tube, a column, or the like before it is exposed to an agent. Any complex that forms on the solid phase is detected at the end of the period of exposure. For example, a protein of the present invention can be anchored to a solid surface, and the test compound (which is

10 not anchored and can be labeled, directly or indirectly) is added to the surface bearing the anchored protein. Un-reacted (e.g., unbound) components can be removed (by, e.g., washing) under conditions that allow any complexes formed to remain immobilized on the solid surface, where they can be detected (e.g., by virtue of a label attached to the protein or the agent or with a labeled antibody

15 that specifically binds an immobilized component and may, itself, be directly or indirectly labeled).

One can immobilize either a protein of the present invention or an antibody to which it specifically binds to facilitate separation of complexed (or bound) protein from uncomplexed (or unbound) protein. Such immobilization

20 can also make it easier to automate the assay, and fusing the proteins of the invention to heterologous proteins can facilitate their immobilization. For example, proteins fused to glutathione-S-transferase can be adsorbed onto glutathione sepharose beads (Sigma Chemical Co., St. Louis, MO) or glutathione derivatized microtiter plates, then combined with the agent and

25 incubated under conditions conducive to complex formation (e.g., conditions in which the salt and pH levels are within physiological levels). Following incubation, the solid phase is washed to remove any unbound components (where the solid phase includes beads, the matrix can be immobilized), the presence or absence of a complex is determined. Alternatively, complexes can

be dissociated from a matrix, and the level of protein binding or activity can be determined using standard techniques.

Immobilization can be achieved with methods known in the art. For example, biotinylated protein can be prepared from biotin-NHS (N-hydroxy-succinimide) using techniques known in the art (e.g., the biotinylation kit from  
5 Pierce Chemicals, Rockford, IL) and immobilized in the wells of streptavidin-coated tissue culture plates (also from Pierce Chemical).

The screening assays of the invention can employ antibodies that react with the proteins of the invention but do not interfere with their activity. These  
10 antibodies can be derivatized to a solid surface, where they will trap a protein of the invention. Any interaction between a protein of the invention and an agent can then be detected using a second antibody that specifically binds the complex formed between the protein of the invention and the agent to which it is bound.

Cell-free assays can also be conducted in a liquid phase, in which case  
15 any reaction product can be separated (and thereby detected) by, for example: differential centrifugation (Rivas and Minton, Trends Biochem Sci 18:284-7, 1993); chromatography (e.g., gel filtration or ion-exchange chromatography); electrophoresis [see, e.g., Ausubel et al., Eds., Current Protocols in Molecular  
20 Biology, J. Wiley & Sons, New York, N.Y., (1999)]; or immunoprecipitation [see, e.g., Ausubel et al. (supra); see also Heegaard, J. Mol. Recognit. 11:141-148, (1998) and Hage and Tweed, J. Chromatogr. Biomed. Sci. Appl. 699:499-525, (1997)]. Fluorescence energy transfer (see above) can also be used, and is convenient because binding can be detected without purifying the complex  
25 from solution. Assays in which the entire reaction of interest is carried out in a liquid phase are sometimes referred to as homogeneous assays.

The screening methods of the invention can also be designed as competition assays in which an agent and a substance that is known to bind a protein of the present invention compete to bind that protein. Depending upon  
30 the order of addition of reaction components and the reaction conditions (e.g.,

whether the reaction is allowed to reach equilibrium), agents that inhibit complex formation can be distinguished from those that disrupt preformed complexes.

In either approach, the order in which reactants are added can be varied to obtain different information about the agents being tested. For example, agents that interfere with the interaction between a gene product and one or more of its binding partners (by, e.g., competing with the binding partner), can be identified by adding the binding partner and the agent to the reaction at about the same time. Agents that disrupt preformed complexes (by, e.g., displacing one of the components from the complex), can be added after a complex containing the gene product and its binding partner has formed.

The proteins of the invention can also be used as "bait proteins" in a two- or three-hybrid assay [see, e.g., U.S. Patent No. 5,283,317; Zervos et al., *Cell* 72:223-232, (1993); Madura et al., *J. Biol. Chem.* 268:12046-12054, (1993); Bartel et al. *Biotechniques* 14:920-924, (1993); Iwabuchi et al., *Oncogene* 8:1693-1696, (1993); and WO 94/10300] to identify other proteins that bind to (e.g., specifically bind to) or otherwise interact with a protein of the invention. Such binding proteins can activate or inhibit the proteins of the invention (and thereby influence the biochemical pathways and events in which those proteins are active).

As noted above, the screening assays of the invention can be used to identify an agent that inhibits the expression of a protein of the invention by, for example, inhibiting the transcription or translation of a nucleic acid that encodes it. In these assays, one can contact a cell or cell free mixture with the agent and then evaluate mRNA or protein expression relative to the levels that are observed in the absence of the agent (a statistically significant increase in expression indicating that the agent stimulates mRNA or protein expression and a decrease (again, one that is statistically significant) indicating tat the agent inhibits mRNA or protein expression). Methods for determining levels of

mRNA or protein expression are known in the art and, here, would employ the nucleic acids, proteins, and antibodies of the present invention.

It should be noted that if desired, two or more of the methods described herein can be practiced together. For example, one can evaluate an agent that  
5 was first identified in a cell-based assay in a cell free assay. Similarly, and the ability of the agent to modulate the activity of a protein of the invention can be confirmed in vivo (e.g., in a transgenic animal).

The screening methods of the present invention can also be used to identify proteins (in the event transcripts of the present invention encode  
10 proteins) that are associated (e.g., causally) with drug resistance. One can then block the activity of these proteins (with, e.g., an antibody of the invention) and thereby improve the ability of a therapeutic agent to exert a desirable effect on a cell or tissue in a subject (e.g., a human patient).

Monitoring the influence of therapeutic agents (e.g., drugs) or other  
15 events (e.g., radiation therapy) on the expression or activity of a biomolecular sequence of the present invention can be useful in clinical trials (a desired extension of the screening assays described above). For example, agents that exert an effect by, in part, altering the expression or activity of a protein of the invention ex vivo can be tested for their ability to do so as the treatment  
20 progresses in a subject. Moreover, in animal or clinical trials, the expression or activity of a nucleic acid can be used, optionally in conjunction with that of other genes, as a "read out" or marker of the phenotype of a particular cell.

### *Detection Assays*

The nucleic acid sequences of the invention can serve as polynucleotide  
25 reagents that are useful in detecting a specific nucleic acid sequence. For example, one can use the nucleic acid sequences of the present invention to map the corresponding genes on a chromosome (and thereby discover which proteins of the invention are associated with genetic disease) or to identify an individual from a biological sample (i.e., to carry out tissue typing, which is  
30 useful in criminal investigations and forensic science). The novel transcripts of

the present invention can be used to identify those tissues or cells affected by a disease (e.g., the nucleic acids of the invention can be used as markers to identify cells, tissues, and specific pathologies, such as cancer), and to identify individuals who may have or be at risk for a particular cancer. Specific methods of detection are described herein and are known to those of ordinary skill in the art.

The nucleic acids of the present invention can be used to determine whether a particular individual is the source of a biological sample (e.g., a blood sample). This is presently achieved by examining restriction fragment length polymorphisms (RFLPs; U.S. Patent No. 5,272,057), and the sequences disclosed here are useful as additional DNA markers for RFLP. For example, one can digest a sample of an individual's genomic DNA, separate the fragments (e.g. by Southern blotting), and expose the fragments to probes generated from the nucleic acids of the present invention (methods employing restriction endonucleases are discussed further below). If the pattern of binding matches that obtained from a tissue of an unknown source, then the individual is the source of the tissue.

The nucleic acids of the present invention can also be used to determine the sequence of selected portions of an individual's genome. For example, the sequences that represent new genes can be used to prepare primers that can be used to amplify an individual's DNA and subsequently sequence it. Panels of DNA sequences (each amplified with a different set of primers) can uniquely identify individuals (as every person will have unique sequences due to allelic differences).

Allelic variation occurs to some degree in the coding regions of these sequences, and to a greater degree in the noncoding regions. Each of the sequences described herein can, to some degree, be used as a standard against which DNA from an individual can be compared for identification purposes. Because greater numbers of polymorphisms occur in the noncoding regions, fewer sequences are necessary to differentiate individuals. The noncoding

sequences disclosed herein can provide positive individual identification with a panel of perhaps 10 to 1,000 primers which each yield a noncoding amplified sequence of 100 bases. If predicted coding sequences are used, a more appropriate number of primers for positive individual identification would be  
5 500-2,000.

If a panel of reagents from the nucleic acids described herein is used to generate a unique identification database for an individual, those same reagents can later be used to identify tissue from that individual. Using the database, the individual, whether still living or dead, can subsequently be linked to even very  
10 small tissue samples.

DNA-based identification techniques, including those in which small samples of DNA are amplified (e.g., by PCR) can also be used in forensic biology. Sequences amplified from tissues (such as hair or skin) or body fluids (such as blood, saliva, or semen) found at a crime scene can be compared to a  
15 standard (e.g., sequences obtained and amplified from a suspect), thereby allowing one to determine whether the suspect is the source of the tissue or bodily fluid.

The nucleic acids of the invention, when used as probes or primers, can target specific loci in the human genome. This will improve the reliability of  
20 DNA-based forensic identifications because the more identifying markers examined, the less likely it is that one individual will be mistaken for another. Moreover, tests that rely on obtaining actual genomic sequence (which is possible here) are more accurate than those in which identification is based on the patterns formed by restriction enzyme generated fragments.

25 The nucleic acids of the invention can also be used to study the expression of the mRNAs in histological sections (i.e., they can be used in situ hybridization). This approach can be useful when forensic pathologists are presented with tissues of unknown origin or when the purity of a population of cells (e.g., a cell line) is in question. The nucleic acids can also be used in  
30 diagnosing a particular condition and in monitoring a treatment regime.



*Predictive Medicine*

The nucleic acids, proteins, antibodies, and cells described hereinabove are generally useful in the field of predictive medicine and, more specifically, are useful in diagnostic and prognostic assays and in monitoring clinical trials.

5 For example, one can determine whether a subject is at risk of developing a disorder associated with a lesion in, or the misexpression of, a nucleic acid of the invention (e.g., a cancer such as pancreatic cancer, breast cancer, or a cancer within the urinary system). In addition, the nucleic acids expressed in tumor tissues and not in normal tissues are markers that can be used to determine  
10 whether a subject has or is likely to develop a particular type of cancer.

The "subject" referred to in the context of any of the methods of the present invention, is a vertebrate animal (e.g., a mammal such as an animal commonly used in experimental studies (e.g. rats, mice, rabbits and guinea pigs); a domesticated animal (e.g., a dog or cat); an animal kept as livestock  
15 (e.g., a pig, cow, sheep, goat, or horse); a non-human primate (e.g. an ape, monkey, or chimpanzee); a human primate; an avian (e.g., a chicken); an amphibian (e.g., a frog); or a reptile. The animal can be an unborn animal (accordingly, the methods of the invention can be used to carry out genetic screening or to make prenatal diagnoses). The subject can also be a human.

20 The methods related to predictive medicine can also be carried out by using a nucleic acid of the invention to, for example detect, in a tissue of a subject: (i) the presence or absence of a mutation that affects the expression of the corresponding gene (e.g., a mutation in the 5' regulatory region of the gene); (ii) the presence or absence of a mutation that alters the structure of the  
25 corresponding gene; (iii) an altered level (i.e., a non-wild type level) of mRNA of the corresponding gene (the proteins of the invention can be similarly used to detect an altered level of protein expression); (iv) a deletion or addition of one or more nucleotides from the nucleic acid sequences of the present invention; (v) a substitution of one or more nucleotides in the nucleic acid sequences of  
30 the present invention (e.g., a point mutation); (vi) a gross chromosomal

rearrangement (e.g., a translocation, inversion, or deletion); or (vii) aberrant modification of a gene corresponding to the nucleic acid sequences of the present invention (e.g., modification of the methylation pattern of the genomic DNA). Similarly, one can test for inappropriate post-translational modification  
5 of any protein encoded. Abnormal expression or abnormal gene or protein structures indicate that the subject is at risk for the associated disorder.

A genetic lesion can be detected by, for example, providing an oligonucleotide probe or primer having a sequence that hybridizes to a sense or antisense strand of a nucleic acid sequence of the present invention, a naturally  
10 occurring mutant thereof, or the 5' or 3' sequences that are naturally associated with the corresponding gene, and exposing the probe or primer to a nucleic acid within a tissue of interest (e.g., a tumor). One can detect hybridization between the probe or primer and the nucleic acid of the tissue by standard methods (e.g., in situ hybridization) and thereby detect the presence or absence of the genetic  
15 lesion. Where the probe or primer specifically hybridizes with a new splice variant, the probe or primer can be used to detect a non-wild type splicing pattern of the mRNA. The antibodies of the invention can be similarly used to detect the presence or absence of a protein encoded by a mutant, mis-expressed, or otherwise deficient gene. Diagnostic and prognostic assays are described  
20 further below.

Qualitative or quantitative analyses (which reveal the presence or absence of a substance or its level of expression or activity, respectively) can be carried out for any one of the nucleic acid sequences of the present invention, or (where the nucleic acid encodes a protein) the proteins they encode, by  
25 obtaining a biological sample from a subject and contacting the sample with an agent capable of specifically binding a nucleic acid represented by the nucleic acid sequences of the present invention or a protein those nucleic acids encode. The conditions in which contacting is performed should allow for specific binding. Suitable conditions are known to those of ordinary skill in the art.  
30 The biological sample can be a tissue, a cell, or a bodily fluid (e.g., blood or

serum), which may or may not be extracted from the subject (i.e., expression can be monitored in vivo).

More specifically, the expression of a nucleic acid sequence can be examined by, for example, Southern or Northern analyses, polymerase chain reaction analyses, or with probe arrays. For example, one can diagnose a condition associated with expression or mis-expression of a gene by isolating mRNA from a cell and contacting the mRNA with a nucleic acid probe with which it can hybridize under stringent conditions (the characteristics of useful probes are known to those of ordinary skill in the art and are discussed elsewhere herein). The mRNA can be immobilized on a surface (e.g., a membrane, such as nitrocellulose or other commercially available membrane) following gel electrophoresis.

Alternatively, one or more nucleic acids (the target sequence or the probe) can be distributed on a two-dimensional array (e.g., a gene chip). Arrays are useful in detecting mutations because a probe positioned on the array can have one or more mismatches to a nucleic acid of the invention (e.g., a destabilizing mismatch). For example, genetic mutations in any of nucleic acid sequences of the present invention can be identified in two-dimensional arrays containing light-generated DNA probes [Cronin et al., Human Mutation 7:244-255, (1996)]. Briefly, when a light-generated DNA probe is used, a first array of probes is used to scan through long stretches of DNA in a sample and a control to identify base changes between the sequences by making linear arrays of sequential overlapping probes. This step allows the identification of point mutations, and it can be followed by use of a second array that allows the characterization of specific mutations by using smaller, specialized probe arrays complementary to all variants or mutations detected. Each mutation array is composed of parallel probe sets, one complementary to the wild-type gene and the other complementary to the mutant gene. Arrays are discussed further below; see also; Kozal et al. [Nature Medicine 2:753-759, (1996)].

The level of an mRNA in a sample can also be evaluated with a nucleic acid amplification technique e.g., RT-PCR (U.S. Patent No. 4,683,202), ligase chain reaction [LCR; Barany, Proc. Natl. Acad. Sci. USA 88:189-193, (1991)]; LCR can be particularly useful for detecting point mutations), self sustained  
5 sequence replication [Guatelli et al., Proc. Natl. Acad. Sci. USA 87:1874-1878, (1990)], transcriptional amplification system [Kwoh et al., Proc. Natl. Acad. Sci. USA 86:1173-1177, (1989)], Q-Beta Replicase [Lizardi et al., Bio/Technology 6:1197, (1988)], or rolling circle replication (U.S. Patent No. 5,854,033). Following amplification, the nucleic acid can be detected using  
10 techniques known in the art. Amplification primers are a pair of nucleic acids that anneal to 5' or 3' regions of a gene (plus and minus strands, respectively, or vice-versa) at some distance (possibly a short distance) from one another. For example, each primer can consist of about 10 to 30 nucleotides and bind to sequences that are about 50 to 200 nucleotides apart. Serial analysis of gene  
15 expression can be used to detect transcript levels (U.S. Patent No. 5,695,937). Other useful amplification techniques (useful in, for example, detecting an alteration in a gene) include anchor PCR or RACE PCR.

Mutations in the gene sequences of the invention can also be identified by examining alterations in restriction enzyme cleavage patterns. For example,  
20 one can isolate DNA from a sample cell or tissue and a control, amplify it (if necessary), digest it with one or more restriction endonucleases, and determine the length(s) of the fragment(s) produced (e.g., by gel electrophoresis). If the size of the fragment obtained from the sample is different from the size of the fragment obtained from the control, there is a mutation in the DNA in the  
25 sample tissue. Sequence specific ribozymes (see, for example, U.S. Patent No. 5,498,531) can be used to detect specific mutations by development or loss of a ribozyme cleavage site.

Any sequencing reaction known in the art (including those that are automated) can also be used to determine whether there is a mutation, and, if so,  
30 how the mutant differs from the wild type sequence. Mutations can also be

identified by using cleavage agents to detect mismatched bases in RNA/RNA or RNA/DNA duplexes [Myers et al., Science 230:1242, (1985); Cotton et al., Proc. Natl. Acad. Sci. USA 85:4397, (1988); Saleeba et al., Methods Enzymol. 217:286-295, (1992)]. Mismatch cleavage reactions employ one or more  
5 proteins that recognize mismatched base pairs in double-stranded DNA (so called "DNA mismatch repair" enzymes; e.g., the mutY enzyme of E. coli cleaves A at G/A mismatches and the thymidine DNA glycosylase from HeLa cells cleaves T at G/T mismatches [see Hsu et al., Carcinogenesis 15:1657-1662, (1994) and U.S. Patent No. 5,459,039].

10 Alterations in electrophoretic mobility can also be used to identify mutations. For example, single strand conformation polymorphism (SSCP) can be used to detect differences in electrophoretic mobility between mutant and wild type nucleic acids [Orita et al., Proc. Natl. Acad. Sci. USA 86:2766, (1989); see also Cotton Mutat. Res. 285:125-144, (1993); and Hayashi, Genet.  
15 Anal. Tech. Appl. 9:73-79, (1992)]. Single-stranded DNA fragments of sample and control nucleic acids are denatured and allowed to renature. The secondary structure of single-stranded nucleic acids varies according to sequence, and the resulting alteration in electrophoretic mobility enables the detection of even a single base change. The sensitivity of the assay is enhanced when RNA (rather  
20 than DNA) is used because RNA's secondary structure is more sensitive to a change in sequence. See also Keen et al., Trends Genet. 7:5, (1991). The movement of mutant or wild-type fragments through gels containing a gradient of denaturant is also informative.

When denaturing gradient gel electrophoresis [DGGE; Myers et al.,  
25 Nature 313:495, (1985)] is used, DNA can be modified so it will not completely denature (this can be done by, for example by adding a GC clamp of approximately 40 bp of high-melting GC-rich DNA by PCR). A temperature gradient can be used in place of a denaturing gradient to identify differences in the mobility of control and sample DNA [Rosenbaum and Reissner, Biophys.  
30 Chem. 265:12753, (1987)].

Point mutations can also be detected by selective oligonucleotide hybridization, selective amplification, or selective primer extension [Point et al., Nature 324:163, (1986); Saiki et al., Proc. Natl. Acad. Sci. USA 86:6230, (1989)] or by chemical ligation of oligonucleotides as described in Xu et al.,  
5 Nature Biotechnol. 19:148, (2001). Allele specific amplification technology can also be used [see, e.g., Gibbs et al., Nucleic Acids Res. 17:2437-2448, (1989); Prossner, Tibtech. 11:238, (1993); and Barany, Proc. Natl. Acad. Sci. USA 88:189, (1991)].

When analysis of a gene or protein is carried out in a cell or tissue  
10 sample, the cell or tissue can be immobilized on a support, typically a glass slide, and then contacted with a probe that can hybridize to the nucleic acid or protein of interest.

The detection methods of the invention can be carried out with appropriate controls (e.g., analyses can be conducted in parallel with a sample  
15 known to contain the target sequence and a target known to lack it).

Various approaches can be used to determine protein expression or activity. For example, one can evaluate the amount of protein in a sample by exposing the sample to an antibody that specifically binds the protein of interest. The antibodies described above (e.g., monoclonal antibodies,  
20 detectably labeled antibodies, intact antibodies and fragments thereof) can be used. The methods can be carried out in-vitro (e.g., one can perform an enzyme linked immunosorbent assay (ELISA), an immunoprecipitation, an immunofluorescence analysis, an enzyme immunoassay (EIA), a radioimmunoassay (RIA), or a Western blot analysis) or in vivo (e.g., one can  
25 introduce a labelled antibody that specifically binds to a protein of the present invention into a subject and then detect it by a standard imaging technique). Alternatively, the sample can be labeled and then contacted with an antibody. For example, one can biotinylate the sample, contact it with an antibody (e.g., an antibody positioned on an antibody array) and then detect the bound sample  
30 (e.g., with avidin coupled to a fluorescent label). As with methods to detect

nucleic acids, appropriate control studies can be performed in parallel with those designed to detect protein expression.

The diagnostic molecules disclosed herein can be assembled as kits. Accordingly, the invention features kits for detecting the presence of the biomolecular sequences of the present invention in a biological sample. The kit can include a probe (e.g., a nucleic acid sequence or an antibody), a standard and, optionally, instructions for use. More specifically, antibody-based kits can include a first antibody (e.g., in solution or attached to a solid support) that specifically binds a protein of the present invention and, optionally, a second, different antibody that specifically binds to the first antibody and is conjugated to a detectable agent. Oligonucleotide-based kits can include an oligonucleotide (e.g., a labeled oligonucleotide) that hybridizes with one of the nucleic acids of the present invention under stringent conditions or a pair of oligonucleotides that can be used to amplify a nucleic acid sequence of the present invention. The kits can also include a buffering agent, a preservative, a protein-stabilizing agent, or a component necessary for detecting any included label (e.g., an enzyme or substrate). The kits can also contain a control sample or a series of control samples that can be assayed and compared to the test sample contained. Each component of the kit can be enclosed within an individual container, and all of the various containers can be within a single package.

The detection methods described herein can identify a subject who has, or is at risk of developing, a disease, disorder, condition, or syndrome (the term "disease" is used to encompass all deviations from a normal state) associated with aberrant or unwanted expression or activity of a biomolecular sequence of the present invention. The detection methods also have prognostic value (e.g., they can be used to determine whether or not it is likely that a subject will respond positively (i.e., be effectively treated with) to an agent (e.g., a nucleic acid, protein, small molecule or other drug)). Samples can also be obtained

from a subject during the course of treatment to monitor the treatment's efficacy at a cellular level.

The present invention also features methods of evaluating a sample by creating a gene expression profile for the sample that includes the level of expression of one or more of biomolecular sequences of the present invention. The sample's profile can be compared with that of a reference profile, either of which can be obtained by the methods described herein (e.g., by obtaining a nucleic acid from the sample and contacting the nucleic acid with those on an array). As with other detection methods, profile-based assays can be performed prior to the onset of symptoms (in which case they can be diagnostic), prior to treatment (in which case they can be predictive) or during the course of treatment (in which case they serve as monitors) [see, e.g., Golub et al., Science 286:531, (1999)].

As described hereinabove, the screening methods of the invention can be used to identify candidate therapeutic agents, and those agents can be evaluated further by examining their ability to alter the expression of one or more of the proteins of the invention. For example, one can obtain a cell from a subject, contact the cell with the agent, and subsequently examine the cell's expression profile with respect to a reference profile (which can be, for example, the profile of a normal cell or that of a cell in a physiologically acceptable condition). The agent is evaluated favorably if the expression profile in the subject's cell is, following exposure to the agent, more similar to that of a normal cell or a cell in a physiologically acceptable condition. A control assay can be performed with, for example, a cell that is not exposed to the agent.

Expression profiles (obtained by evaluating either nucleic acid or protein expression) are also useful in evaluating subjects. One can obtain a sample from a subject (either directly or indirectly from a caregiver), create an expression profile, and, optionally, compare the subject's expression profile to one or more reference profiles and/or select a reference profile most similar to that of the subject. A variety of routine statistical measures can be used to



compare two reference profiles. One possible metric is the length of the distance vector that is the difference between the two profiles. Each of the subject and reference profile is represented as a multi-dimensional vector, wherein each dimension is a value in the profile.

5       The result, which can be communicated to the subject, a caregiver, or another interested party, can be the subject's expression profile per se, a result of a comparison of the subject's expression profile with another profile, a most similar reference profile, or a descriptor of any of these. Communication can be mediated by a computer network (e.g., in the form of a computer transmission  
10       such as a computer data signal embedded in a carrier wave).

Accordingly, the invention also features a computer medium having executable code for effecting the following steps: receive a subject expression profile; access a database of reference expression profiles; and either i) select a matching reference profile most similar to the subject expression profile, or ii)  
15       determine at least one comparison score for the similarity of the subject expression profile to at least one reference profile. The subject expression profile and the reference expression profile each include a value representing the level of expression of one or more of the biomolecular sequences of the present invention.

#### 20                               *Arrays and uses thereof*

The present invention also encompasses arrays that include a substrate having a plurality of addresses, at least one of which includes a capture probe that specifically binds or hybridizes to a nucleic acid represented by any one of the biomolecular sequences of the present invention. The array can have a  
25       density of at least 10, 50, 100, 200, 500, 1,000, 2,000, or 10,000 or more addresses/cm<sup>2</sup>, or densities between these. In some embodiments, the plurality of addresses includes at least 10, 100, 500, 1,000, 5,000, 10,000, or 50,000 addresses, while in other embodiments, the plurality of addresses can be equal to, or less than, those numbers.

Regardless of whether the array contains nucleic acids (as probes or targets) or proteins (as probes or targets), the substrate can be two-dimensional (formed, e.g., by a glass slide, a wafer (e.g., silica or plastic), or a mass spectroscopy plate) or three-dimensional (formed, e.g., by a gel or pad).

5 Addresses in addition to the addresses of the plurality can be disposed on the array.

At least one address of the plurality can include a nucleic acid capture probe that hybridizes specifically to one or more of the nucleic acid sequences of the present invention. In certain embodiments, a subset of addresses of the plurality will be occupied by a nucleic acid capture probe for one of the nucleic acid sequences of the present invention; each address in the subset can bear a capture probe that hybridizes to a different region of a selected nucleic acid. In other embodiments, the probe at each address is unique, overlapping, and complementary to a different variant of a selected nucleic acid (e.g., an allelic variant, or all possible hypothetical variants). If desired, the array can be used to sequence the selected nucleic acid by hybridization (see, e.g., U.S. Patent No. 5,695,940). Alternatively, the capture probe can be a protein that specifically binds to a protein of the present invention or a fragment thereof (e.g., a naturally-occurring interaction partners of a protein of the invention or an antibody described herein). In some instances (e.g., in the event of an autoimmune disease), it is significant that a subject produces antibodies, and the arrays described herein can be used to detect those antibodies. More generally, an array that contains some or all of the proteins of the present invention can be used to detect any substance to which one or more those proteins bind (e.g., a natural binding partner, an antibody, or a synthetic molecule).

10  
15  
20  
25

An array can be generated by methods known to those of ordinary skill in the art. For example, an array can be generated by photolithographic methods (see, e.g., U.S. Patent Nos. 5,143,854; 5,510,270; and 5,527,681), mechanical methods (e.g., directed-flow methods as described in U.S. Patent No. 5,384,261), pin-based methods (e.g., as described in U.S. Pat. No.

30

5,288,514), and bead-based techniques (e.g., as described in PCT US/93/04145). Methods of producing protein-based arrays are described in, for example, De Wildt et al. [Nature Biotech. 18:89-994, (2000)], Lueking et al. [Anal. Biochem. 270:103-111, (1999)], Ge [Nucleic Acids Res. 28:e3, I-VII, (2000)], MacBeath and Schreiber [Science 289:1760-1763, (2000)], and WO 99/51773A1. Addresses in addition to the address of the plurality can be disposed on the array.

The arrays described above can be used to analyze the expression of any of the biomolecular sequences of the present invention. For example, one can contact an array with a sample and detect binding between a component of the sample and a component of the array. In the event nucleic acids are analyzed, one can amplify the nucleic acids obtained from a sample prior to their application to the array. The array can also be used to examine tissue-specific gene expression. For example, the nucleic acids or proteins of the invention (all or a subset thereof) can be distributed on an array that is then exposed to nucleic acids or proteins obtained from a particular tissue, tumor, or cell type. If a sufficient number of diverse samples are analyzed, clustering (e.g., hierarchical clustering, k-means clustering, Bayesian clustering and the like) can be used to identify other genes that are co-regulated with those of the invention. The array can be used not only to determine tissue specific expression, but also to ascertain the level of expression of a battery of genes.

Array analysis of the nucleic acids or proteins of the invention can be used to study the effect of cell-cell interactions or therapeutic agents on the expression of those nucleic acids or proteins. For example, nucleic acid or protein that has been obtained from a cell that has been placed in the vicinity of a tissue that has been perturbed in some way can be obtained and exposed to the probes of an array. Thus, one can use the methods of the invention to determine the effect of one cell type on another (i.e., the response (e.g., a change in the type or quantity of nucleic acids or proteins expressed) to a biological stimulus can be determined). Similarly, nucleic acid or protein that has been obtained

from a cell that has been treated with an agent can be obtained and exposed to the probes of an array. In this scenario, one can determine how the therapeutic agent affects the expression of any of the biomolecular sequences of the present invention. Appropriate controls (e.g., assays using cells that have not received a biological stimulus or a potentially therapeutic treatment) can be performed in parallel. Moreover, desirable and undesirable responses can be detected. If an event (e.g., exposure to a biological stimulus or therapeutic compound) has an undesirable effect on a cell, one can either avoid the event (by, e.g., prescribing an alternative therapy) or take steps to counteract or neutralize it.

10 In more straightforward assays, the arrays described here can be used to monitor the expression of one or more of the biomolecular sequences of the present invention, with respect to time. Such analyses allow one to characterize a disease process associated with the examined sequence.

The arrays are also useful for ascertaining the effect of the expression of a gene on the expression of other genes in the same cell or in different cells (e.g., ascertaining the effect of the expression of any one of the biomolecular sequences of the present invention on the expression of other genes). If altering the expression of one gene has a deleterious effect on the cell (due to its effect on the expression of other genes) one can, again, avoid that effect (by, e.g., selecting an alternate molecular target or counteracting or neutralizing the effect).

### *Markers*

The molecules of the present invention are also useful as markers of: (i) a cell or tissue type; (ii) disease; (iii) a pre-disease state; (iv) drug activity, and (v) predisposition for disease.

Using the methods described herein, the presence or amount of the biomolecular sequences of the present invention, can be detected and correlated with one or more biological states (e.g., a disease state or a developmental state). When used in this way, the compositions of the invention serve as surrogate markers; they provide an objective indicia of the presence or extent of

a disease (e.g., cancer). Surrogate markers are particularly useful when a disease is difficult to assess with standard methods (e.g., when a subject has a small tumor or when pre-cancerous cells are present). It follows that surrogate markers can be used to assess a disease before a potentially dangerous clinical endpoint is reached. Other examples of surrogate markers are known in the art (see, e.g., Koomen et al., *J. Mass Spectrom.* 35:258-264, 2000, and James, *AIDS Treatment News Archive* 209, 1994).

The biomolecular sequences of the present invention, can also serve as pharmacodynamic markers, which provide an indicia of a therapeutic result. As pharmacodynamic markers are not directly related to the disease for which the drug is being administered, their presence (or levels of expression) indicates the presence or activity of a drug in a subject (i.e., the pharmacodynamic marker may indicate the concentration of a drug in a biological tissue, as the gene or protein serving as the marker is either expressed or transcribed (or not) in the body in relationship to the level or activity of the drug). One can also monitor the distribution of a drug with a pharmacodynamic marker (e.g., these markers can be used to determine whether a drug is taken up by a particular cell type). The presence or amount of pharmacodynamic markers can be related to the drug per se or to a metabolite produced from the drug. Thus, these markers can indicate the rate at which a drug is broken down in vivo. Pharmacodynamic markers can be particularly sensitive (e.g., even a small amount of a drug may activate substantial transcription or translation of a marker), and they are therefore useful in assessing drugs that are administered at low doses. Examples regarding the use of pharmacodynamic markers are known in the art and include: U.S. Patent No. 6,033,862; Hattis et al. *Env. Health Perspect.* 90: 229-238, (1991); Schentag, *Am. J. Health-Syst. Pharm.* 56 Suppl. 3:S21-S24, (1999); and Nicolau, *Am. J. Health-Syst. Pharm.* 56 Suppl. 3: S16-S20, (1991).

The biomolecular sequences of the present invention, are also useful as pharmacogenomic markers, which can provide an objective correlate to a specific clinical drug response or susceptibility in a particular subject or class of

subjects [see, e.g., McLeod et al., *Eur. J. Cancer* 35:1650-1652, (1999)]. The presence or amount of the pharmacogenomic marker is related to the predicted response of a subject to a specific drug (or type of drug) prior to administration of the drug. By assessing one or more pharmacogenomic markers in a subject, the drug therapy that is most appropriate for the subject, or which is predicted to have a greater likelihood of success, can be selected. For example, based on the presence or amount of RNA or protein associated with a specific tumor marker in a subject, an optimal drug or treatment regime can be prescribed for the subject.

More generally, pharmacogenomics addresses the relationship between an individual's genotype and that individual's response to a foreign compound or drug. Differences in the way individual subjects metabolize therapeutics can lead to severe toxicity or therapeutic failure because metabolism alters the relation between dose and blood concentration of the pharmacologically active drug. Thus, a physician would consider the results of pharmacogenomic studies when determining whether to administer a composition of the present invention and how to tailor a therapeutic regimen for the subject.

Pharmacogenomics deals with clinically significant hereditary variations in the response to drugs due to altered drug disposition and abnormal action in affected persons. See, e.g., Eichelbaum et al., *Clin. Exp. Pharmacol. Physiol.* 23:983-985, (1996), and Linder et al., *Clin. Chem.* 43:254-266, (1997). In general, two types of pharmacogenetic conditions can be differentiated. Genetic conditions transmitted as a single factor can: (i) alter the way drugs act on the body (altered drug action) or (ii) the way the body acts on drugs (altered drug metabolism). These pharmacogenetic conditions can occur either as rare genetic defects or as naturally-occurring polymorphisms.

One approach that can be used to identify genes that predict drug response, known as "a genome-wide association," relies primarily on a high-resolution map of the human genome consisting of already known gene-related markers (e.g., a "bi-allelic" gene marker map that consists of 60,000-100,000

polymorphic or variable sites on the human genome, each of which has two variants.) Such a high-resolution genetic map can be compared to a map of the genome of each of a statistically significant number of patients taking part in a Phase II/III drug trial to identify markers associated with a particular observed drug response or side effect. Alternatively, a high resolution map can be generated from a combination of some ten-million known single nucleotide polymorphisms (SNPs; a common alteration that occurs in a single nucleotide base in a stretch of DNA) in the human genome. For example, a SNP may occur once per every 1000 bases of DNA. While a SNP may be involved in a disease process, the vast majority may not be disease-associated. Given a genetic map based on the occurrence of such SNPs, individuals can be grouped into genetic categories depending on a particular pattern of SNPs in their individual genome. In such a manner, treatment regimens can be tailored to groups of genetically similar individuals, taking into account traits that may be common among such genetically similar individuals.

Two alternative methods, the "candidate gene approach" and "gene expression profiling," can be used to identify pharmacogenomic markers. According to the first method, if a gene that encodes a drug's target is known, all common variants of that gene can be fairly easily identified in the population, and one can determine whether having one version of the gene versus another is associated with a particular drug response. In the second approach, the gene expression of an animal dosed with a drug (e.g., a composition of the invention) can reveal whether gene pathways related to toxicity have been activated.

Information generated using one or more of the approaches described above can be used in designing therapeutic or prophylactic treatments that are less likely to fail or to produce adverse side effects when a subject is treated with a therapeutic composition.

*Informatics*

The biomolecular sequences of the present invention can be provided in a variety of media to facilitate their use. For example, one or more of the sequences (e.g., subsets of the sequences expressed in a defined tissue type) can be provided as a manufacture (e.g., a computer-readable storage medium such as a magnetic, optical, optico-magnetic, chemical or mechanical information storage device). The manufacture can provide a nucleic acid or amino acid sequence in a form that will allow examination of the manufacture in ways that are not applicable to a sequence that exists in nature or in purified form. The sequence information can include full-length sequences, fragments thereof, polymorphic sequences including single nucleotide polymorphisms (SNPs), epitope sequence, and the like.

The computer readable storage medium further includes sequence annotations (as described in Example 10 of the Examples section).

The computer readable storage medium can further include information pertaining to generation of the data and/or potential uses thereof.

As used herein, a "computer-readable medium" refers to any medium that can be read and accessed directly by a machine [e.g., a digital or analog computer; e.g., a desktop PC, laptop, mainframe, server (e.g., a web server, network server, or server farm), a handheld digital assistant, pager, mobile telephone, or the like]. Computer-readable media include: magnetic storage media, such as floppy discs, hard disc storage medium, and magnetic tape; optical storage media such as CD-ROM; electrical storage media such as RAM, ROM, EPROM, EEPROM, flash memory, and the like; and hybrids of these categories such as magnetic/optical storage media.

A variety of data storage structures are available to those of ordinary skill in the art and can be used to create a computer-readable medium that has recorded one or more (or all) of the nucleic acids and/or amino acid sequences of the present invention. The data storage structure will generally depend on the means chosen to access the stored information. In addition, a variety of data



processor programs and formats can be used to store the sequence information of the present invention on machine or computer-readable medium. The sequence information can be represented in a word processing text file, formatted in commercially-available software such as WordPerfect and Microsoft Word, or represented in the form of an ASCII file, stored in a database application, such as DB2, Sybase, Oracle, or the like. One of ordinary skill in the art can readily adapt any number of data processor structuring formats (e.g., text file or database) to obtain machine or computer-readable medium having recorded thereon the sequence information of the present invention.

The sequence information and annotations are stored in a relational database (such as Sybase or Oracle) that can have a first table for storing sequence (nucleic acid and/or amino acid sequence) information. The sequence information can be stored in one field (e.g., a first column) of a table row and an identifier for the sequence can be stored in another field (e.g., a second column) of the table row. The database can have a second table (to, for example, store annotations). The second table can have a field for the sequence identifier, a field for a descriptor or annotation text (e.g., the descriptor can refer to a functionality of the sequence), a field for the initial position in the sequence to which the annotation refers, and a field for the ultimate position in the sequence to which the annotation refers. Examples for annotation to nucleic acid sequences and amino acid sequences are provided in Examples 10 and 14-20 of the Examples section.

### *Pharmaceutical Compositions*

The nucleic acids, fragments thereof, hybrid sequences of which they are a part, and gene constructs containing them; proteins, fragments thereof, chimeras, and antibodies that specifically bind thereto; and cells, including those that are engineered to express the nucleic acids or proteins of the invention) can be incorporated into pharmaceutical compositions. These compositions typically also include a solvent, a dispersion medium, a coating,

an antimicrobial (e.g., an antibacterial or antifungal) agent, an absorption delaying agent (when desired, such as aluminum monostearate and gelatin), or the like, compatible with pharmaceutical administration (see below). Active compounds, in addition to those of the present invention, can also be included  
5 in the composition and may enhance or supplement the activity of the present agents.

The composition will be formulated in accordance with their intended route of administration. Acceptable routes include oral or parenteral routes (e.g., intravenous, intradermal, transdermal (e.g., subcutaneous or topical), or  
10 transmucosal (i.e., across a membrane that lines the respiratory or anogenital tract). The compositions can be formulated as a solution or suspension and, thus, can include a sterile diluent (e.g., water, saline solution, a fixed oil, polyethylene glycol, glycerine, propylene glycol or another synthetic solvent); an antimicrobial agent (e.g., benzyl alcohol or methyl parabens; chlorobutanol,  
15 phenol, ascorbic acid, thimerosal, and the like); an antioxidant (e.g., ascorbic acid or sodium bisulfite); a chelating agent (e.g., ethylenediaminetetraacetic acid); or a buffer (e.g., an acetate-, citrate-, or phosphate-based buffer). When necessary, the pH of the solution or suspension can be adjusted with an acid (e.g., hydrochloric acid) or a base (e.g., sodium hydroxide). Proper fluidity  
20 (which can ease passage through a needle) can be maintained by a coating such as lecithin, by maintaining the required particle size (in the case of a dispersion), or by the use of surfactants.

The compositions of the invention can be prepared as sterile powders (by, e.g., vacuum drying or freeze-drying), which can contain the active  
25 ingrediaent plus any additional desired ingredient from a previously sterile-filtered solution.

Oral compositions generally include an inert diluent or an edible carrier. For example, the active compound can be incorporated with excipients and used in the form of tablets, troches, or capsules (e.g., gelatin capsules). Oral  
30 compositions can be prepared using fluid carries and used as mouthwashes.

The tablets etc. can also contain a binder (e.g., microcrystalline cellulose, gum tragacanth, or gelatin); an excipient (e.g., starch or lactose), a disintegrating agent (e.g., alginic acid, Primogel, or corn starch); a lubricant (e.g., magnesium stearate or Sterotes); a glidant (e.g., colloidal silicon dioxide); a sweetening agent (e.g., sucrose or saccharine); or a flavoring agent (e.g., peppermint, methyl salicylate, or orange flavoring).

For administration by way of the respiratory system, the compositions can be formulated as aerosol sprays (e.g., from a pressured container or dispenser that contains a suitable propellant (e.g., a gas such as carbon dioxide), or a nebulizer. The ability of a composition to cross a biological barrier can be enhanced by agents known in the art. For example, detergents, bile salts, and fusidic acid derivatives can facilitate transport across the mucosa (and therefore, be included in nasal sprays or suppositories).

For topical administration, the active compounds are formulated into ointments, salves, gels, or creams according to methods known in the art.

Controlled release can also be achieved by using implants and microencapsulated delivery systems (see, e.g., the materials commercially available from Alza Corporation and Nova Pharmaceuticals, Inc.; see also U.S. Patent No. 4,522,811 for the use of liposome-based suspensions).

The pharmaceutical compositions of the invention can be formulated in dosage units (i.e., physically discrete units containing a predetermined quantity of the active compound) for uniformity and ease of administration.

The toxicity and therapeutic efficacy of any given compound can be determined by standard pharmaceutical procedures carried out in cell culture or in experimental animals. For example, one of ordinary skill in the art can routinely determine the LD50 (the dose lethal to 50% of the population) and the ED50 (the dose therapeutically effective in 50% of the population). The dose ratio between toxic and therapeutic effects is the therapeutic index. Compounds that exhibit high therapeutic indices are preferred. While compounds that exhibit toxic side effects may be used, care should be taken to design a delivery

system that targets such compounds to the site of affected tissue in order to minimize potential damage to uninfected cells and, thereby, reduce side effects.

The data obtained from the cell culture assays and animal studies described hereinabove can be used to formulate a range of dosage for use in humans (preferably a dosage within a range of circulating concentrations that include the ED50 with little or no toxicity). The dosage may vary within this range depending upon the formulation and the route of administration. For any compound used in the method of the invention, the therapeutically effective dose can be estimated initially from cell culture assays. A dose may be formulated in animal models to achieve a circulating plasma concentration range that includes the IC50 (i.e., the concentration of the test compound which achieves a half-maximal inhibition of symptoms) as determined in cell culture. Such information can be used to more accurately determine useful doses in humans. Levels in plasma may be measured, for example, by high performance liquid chromatography.

A therapeutically effective amount of a protein of the present invention can range from about 0.001 to 30 mg/kg body weight (e.g., about 0.01 to 25 mg/kg, about 0.1 to 20 mg/kg, or about 1 to 10 (e.g., 2-9, 3-8, 4-7, or 5-6) mg/kg). The protein can be administered one time per week for between about 1 to 10 weeks (e.g., 2 to 8 weeks, 3 to 7 weeks, or about 4, 5, or 6 weeks). However, a single administration can also be efficacious. Certain factors can influence the dosage and timing required to effectively treat a subject. These factors include the severity of the disease, previous treatments, and the general health or age of the subject.

When the active ingredient is an antibody, the dosage can be about 0.1 mg/kg of body weight (generally 10-20 mg/kg). If the antibody is to act in the brain, a dosage of 50 mg/kg to 100 mg/kg is usually appropriate. Generally, partially human antibodies and fully human antibodies have a longer half-life within the human body than other antibodies. Accordingly, lower dosages and less frequent administration are often possible with these types of antibodies.

Modifications such as lipidation can be used to stabilize antibodies and to enhance uptake and tissue penetration [e.g., into the brain; see Cruikshank et al., J. Acquired Immune Deficiency Syndromes and Human Retrovirology 14:193, (1997)].

5 As noted above, the present invention encompasses agents (e.g., small molecules) that modulate expression or activity of a nucleic acid represented by any of biomolecular sequences of the present invention. Exemplary doses of these agents include milligram or microgram amounts of the small molecule per kilogram of subject or sample weight (e.g., about 1-500 mg/kg; about 100  
10 mg/kg; about 5 mg/kg; about 1 mg/kg; or about 50 µg/kg). Appropriate doses of a small molecule depend upon the potency of the small molecule with respect to the expression or activity to be modulated. When one or more of these small molecules is to be administered to an animal (e.g., a human) to modulate expression or activity of nucleic acid or protein of the invention, a physician,  
15 veterinarian, or researcher may prescribe a relatively low dose at first, subsequently increasing the dose until an appropriate response is obtained. In addition, it is understood that the specific dose level for any particular animal subject will depend upon a variety of factors including the activity of the specific compound employed, the age, body weight, general health, gender, and  
20 diet of the subject, the time of administration, the route of administration, the rate of excretion, any drug combination, and the degree of expression or activity to be modulated.

Pharmaceutical compositions of the present invention may also include a therapeutic moiety such as a cytotoxin (i.e., an agent that is detrimental to a  
25 cell), a therapeutic agent, or a radioactive ion can be conjugated to the biomolecular sequences of the present invention or related compositions, described hereinabove (e.g., antibodies, antisense molecules, ribozymes etc.). The cytotoxin can be, for example, taxol, cytochalasin B, gramicidin D, ethidium bromide, emetine, mitomycin, etoposide, tenoposide, vincristine,  
30 vinblastine, colchicin, doxorubicin, daunorubicin, dihydroxy anthracin dione,

mitoxantrone, mithramycin, actinomycin D, 1-dehydrotestosterone, glucocorticoids, procaine, tetracaine, lidocaine, propranolol, puromycin, maytansinoids (e.g., maytansinol; see U.S. Patent No. 5,208,020), CC-1065 (see U.S. Patent Nos. 5,475,092, 5,585,499, and 5,846,545) and analogs or  
5 homologs thereof. Therapeutic agents include antimetabolites (e.g., methotrexate, 6-mercaptopurine, 6-thioguanine, cytarabine, 5-fluorouracil decarbazine), alkylating agents (e.g., mechlorethamine, thioepa chlorambucil, CC-1065, melphalan, carmustine (BSNU) and lomustine (CCNU), cyclophosphamide, busulfan, dibromomannitol, streptozotocin, mitomycin C,  
10 and cis-dichlorodiamine platinum (II) (DDP) cisplatin), anthracyclines (e.g., daunorubicin (formerly daunomycin) and doxorubicin), antibiotics (e.g., dactinomycin (formerly actinomycin), bleomycin, mithramycin, and anthramycin (AMC)), and anti-mitotic agents (e.g., vincristine, vinblastine, taxol and maytansinoids). Radioactive ions include, but are not limited to  
15 iodine, yttrium and praseodymium.

Other therapeutic moieties include, but are not limited to, toxins such as abrin, ricin A, pseudomonas exotoxin, or diphtheria toxin; a protein such as tumor necrosis factor,  $\gamma$ -interferon,  $\beta$ -interferon, nerve growth factor, platelet derived growth factor, tissue plasminogen activator; or, biological response  
20 modifiers such as, for example, lymphokines, interleukin-1 (IL-1), interleukin-2 (IL-2), interleukin-6 (IL-6), granulocyte macrophase colony stimulating factor (GM-CSF), granulocyte colony stimulating factor (G-CSF), or other growth factors.

The nucleic acid molecules of the invention can be inserted into vectors  
25 and used as gene therapy vectors. Gene therapy vectors can be delivered to a subject by, for example, intravenous injection, local administration (see U.S. Patent 5,328,470) or by stereotactic injection (see e.g., Chen et al., Proc. Natl. Acad. Sci. USA 91:3054-3057, 1994). The pharmaceutical preparation of the gene therapy vector can include the gene therapy vector in an acceptable  
30 diluent, or can comprise a slow release matrix in which the gene delivery

vehicle is imbedded. Alternatively, where the complete gene delivery vector can be produced intact from recombinant cells (e.g. retroviral vectors), the pharmaceutical preparation can include one or more cells which produce the gene delivery system. The pharmaceutical compositions of the invention can be included in a container, pack, or dispenser together with instructions for administration.

### *Methods of Treatment*

The present invention provides for both prophylactic and therapeutic methods of treating a subject at risk of (or susceptible to) a disorder or having a disorder associated with aberrant or unwanted expression or activity of a nucleic acid or protein of the invention. "Treatment" encompasses the application or administration of a therapeutic agent to a patient, or to an isolated tissue or cell line (e.g., one obtained from the patient to be treated), with the purpose of curing or lessening the severity of the disease or a symptom associated with the disease.

Whether carried out prophylactically or therapeutically, the methods of the invention can be specifically tailored or modified, based on knowledge obtained from the field of pharmacogenomics (see above).

Thus, the invention provides a method for preventing in a subject, a disease associated with mis-expression of a nucleic acid or protein of the present invention. Such diseases include cellular proliferative and/or differentiative disorders, disorders associated with bone metabolism, immune disorders, cardiovascular disorders, liver disorders, viral diseases, pain or metabolic disorders.

Examples of cellular proliferative and/or differentiative disorders include cancer (e.g., carcinoma, sarcoma, metastatic disorders or hematopoietic neoplastic disorders such as leukemias and lymphomas). A metastatic tumor can arise from a multitude of primary tumor types, including but not limited to those of prostate, colon, lung, breast or liver.

The terms "cancer," "hyperproliferative," and "neoplastic," are used in reference to cells that have exhibited a capacity for autonomous growth (i.e., an abnormal state or condition characterized by rapid cellular proliferation). Hyperproliferative and neoplastic disease states can be categorized as pathologic (i.e., characterizing or constituting a disease state), or can be categorized as non-pathologic (i.e., deviating from normal but not associated with a disease state). The term is meant to include all types of cancerous growths or oncogenic processes, metastatic tissues or malignantly transformed cells, tissues, or organs, irrespective of histopathologic type or stage of invasiveness. "Pathologic hyperproliferative" cells occur in disease states characterized by malignant tumor growth. Examples of non-pathologic hyperproliferative cells include proliferation of cells associated with wound repair.

The terms "cancer" or "neoplasms" include malignancies of the various organ systems, such as affecting lung, breast, thyroid, lymphoid, gastrointestinal, and genito-urinary tract, as well as adenocarcinomas, which include malignancies such as most colon cancers, renal-cell carcinoma, prostate cancer and/or testicular tumors, non-small cell carcinoma of the lung, cancer of the small intestine and cancer of the esophagus.

The term "carcinoma" refers to malignancies of epithelial or endocrine tissues including respiratory system carcinomas, gastrointestinal system carcinomas, genitourinary system carcinomas, testicular carcinomas, breast carcinomas, prostatic carcinomas, endocrine system carcinomas, and melanomas. Exemplary carcinomas include those forming from tissue of the cervix, lung, prostate, breast, head and neck, colon and ovary. The term also includes carcinosarcomas (e.g., which include malignant tumors composed of carcinomatous and sarcomatous tissues). An "adenocarcinoma" refers to a carcinoma derived from glandular tissue or in which the tumor cells form recognizable glandular structures. The term "sarcoma" is art recognized and refers to malignant tumors of mesenchymal derivation. As used herein, the



term "hematopoietic neoplastic disorder(s)" includes diseases involving hyperplastic/neoplastic cells of hematopoietic origin. A hematopoietic neoplastic disorder can arise from myeloid, lymphoid or erythroid lineages, or precursor cells thereof. Preferably, the diseases arise from poorly differentiated acute leukemias (e.g., erythroblastic leukemia and acute megakaryoblastic leukemia). Additional exemplary myeloid disorders include, but are not limited to, acute promyeloid leukemia (APML), acute myelogenous leukemia (AML) and chronic myelogenous leukemia (CML) (see Vaickus, Crit Rev. in Oncol./Hematol. 11:267-97, 1991); lymphoid malignancies include, but are not limited to acute lymphoblastic leukemia (ALL) which includes B-lineage ALL and T-lineage ALL, chronic lymphocytic leukemia (CLL), prolymphocytic leukemia (PLL), hairy cell leukemia (HLL) and Waldenstrom's macroglobulinemia (WM). Additional forms of malignant lymphomas include, but are not limited to non-Hodgkin lymphoma and variants thereof, peripheral T cell lymphomas, adult T cell leukemia/lymphoma (ATL), cutaneous T-cell lymphoma (CTCL), large granular lymphocytic leukemia (LGF), Hodgkin's disease and Reed-Sternberg disease.

The leukemias, including B-lymphoid leukemias, T-lymphoid leukemias, undifferentiated leukemias, erythroleukemia, megakaryoblastic leukemia, and monocytic leukemias are encompassed with and without differentiation; chronic and acute lymphoblastic leukemia, chronic and acute lymphocytic leukemia, chronic and acute myelogenous leukemia, lymphoma, myelo dysplastic syndrome, chronic and acute myeloid leukemia, myelomonocytic leukemia; chronic and acute myeloblastic leukemia, chronic and acute myelogenous leukemia, chronic and acute promyelocytic leukemia, chronic and acute myelocytic leukemia, hematologic malignancies of monocyte-macrophage lineage, such as juvenile chronic myelogenous leukemia; secondary AML, antecedent hematological disorder; refractory anemia; aplastic anemia; reactive cutaneous angioendotheliomatosis; fibrosing disorders involving altered expression in dendritic cells, disorders including systemic sclerosis, E-M

syndrome, epidemic toxic oil syndrome, eosinophilic fasciitis localized forms of scleroderma, keloid, and fibrosing colonopathy; angiomatoid malignant fibrous histiocytoma; carcinoma, including primary head and neck squamous cell carcinoma; sarcoma, including kaposi's sarcoma; fibroadenoma and phyllodes tumors, including mammary fibroadenoma; stromal tumors; phyllodes tumors, including histiocytoma; erythroblastosis; and neurofibromatosis.

Examples of disorders involving the heart or "cardiovascular disorders" include, but are not limited to, a disease, disorder, or state involving the cardiovascular system, e.g., the heart, the blood vessels, and/or the blood. A cardiovascular disorder can be caused by an imbalance in arterial pressure, a malfunction of the heart, or an occlusion of a blood vessel, e.g., by a thrombus. Examples of such disorders include hypertension, atherosclerosis, coronary artery spasm, congestive heart failure, coronary artery disease, valvular disease, arrhythmias, and cardiomyopathies.

As discussed, diseases associated (e.g., causally associated) with over expression of a nucleic acid of the invention (as determined, for example, by the in vivo or ex vivo analyses described above), can be treated with techniques in which one inhibits the expression or activity of the nucleic acid or its gene products. For example, a compound (e.g., an agent identified using an assay described above) that exhibits negative modulatory activity with respect to a nucleic acid of the invention (the expression or over expression of which is causally associated with a disease) can be used to prevent and/or ameliorate that disease or one or more of the symptoms associated with it. The compound can be a peptide, phosphopeptide, small organic or inorganic molecule, or antibody (e.g., a polyclonal, monoclonal, humanized, anti-idiotypic, chimeric or single chain antibodies, and Fab, F(ab')<sub>2</sub> and Fab expression library fragments, scFV molecules, and epitope-binding fragments thereof).

Further, antisense, ribozyme, and triple-helix molecules (see above) that inhibit expression of the target gene (e.g., a gene of the invention) can also be used to reduce the level of target gene expression, thus effectively reducing the

level of target gene activity. If necessary, to achieve a desirable level of gene expression, molecules that inhibit gene expression can be administered with nucleic acid molecules that encode and express target gene polypeptides exhibiting normal target gene activity. Of course, where the assays of the invention indicate that expression or over expression is desirable, the nucleic acid can be introduced into cells via gene therapy methods with little or no treatment with inhibitory agents (this can be done to combat not only under expression, but over secretion of a gene product).

Aptamer molecules (nucleic acid molecules having a tertiary structure that permits them to specifically bind to protein ligands; [see, e.g., Osborne et al., *Curr. Opin. Chem. Biol.* 1: 5-9, (1997) and Patel *Curr. Opin. Chem. Biol.* 1:32-46, (1997)] are also useful therapeutics. Since nucleic acid molecules can usually be more conveniently introduced into target cells than therapeutic proteins may be, aptamers offer a method by which protein activity can be specifically decreased without the introduction of drugs or other molecules that may have pluripotent effects.

As noted above, the nucleic acids of the invention and the proteins they encode can be used as immunotherapeutic agents (to, e.g., elicit an immune response against a protein of interest). However, in some circumstances, undesirable effects occur when a subject is injected with a protein or an epitope that stimulate antibody production. In those circumstances, one can instead generate an immune response with an anti-idiotypic antibody [see, e.g., Herlyn, *Ann. Med.* 31:66-78, 1991 and Bhattacharya-Chatterjee and Foon, *Cancer Treat. Res.* 94:51-68, (1998)]. Effective anti-idiotypic antibodies stimulate the production of anti-anti-idiotypic antibodies, which specifically bind the protein in question. Vaccines directed to a disease characterized by expression of the nucleic acids of the present invention can also be generated in this fashion. In other circumstances, the target antigen is intracellular. In these circumstances, antibodies (including fragments, single chain antibodies, or other types of antibodies described above) can be internalized within a cell by delivering them

with, for example, a lipid-based delivery system (e.g., Lipofectin™ or liposomes). Single chain antibodies can also be administered by delivering nucleotide sequences that encode them to the target cell population (see, e.g., Marasco et al., Proc. Natl. Acad. Sci. USA 90:7889-7893, 1993).

5

Additional objects, advantages, and novel features of the present invention will become apparent to one ordinarily skilled in the art upon examination of the following examples, which are not intended to be limiting. Additionally, each of the various embodiments and aspects of the present invention as delineated hereinabove and as claimed in the claims section below  
10 finds experimental support in the following examples.

### EXAMPLES

Reference is now made to the following examples, which together with  
15 the above descriptions, illustrate the invention in a non limiting fashion.

Generally, the nomenclature used herein and the laboratory procedures utilized in the present invention include molecular, biochemical, microbiological and recombinant DNA techniques. Such techniques are thoroughly explained in the literature. See, for example, "Molecular Cloning: A  
20 laboratory Manual" Sambrook et al., (1989); "Current Protocols in Molecular Biology" Volumes I-III Ausubel, R. M., ed. (1994); Ausubel et al., "Current Protocols in Molecular Biology", John Wiley and Sons, Baltimore, Maryland (1989); Perbal, "A Practical Guide to Molecular Cloning", John Wiley & Sons, New York (1988); Watson et al., "Recombinant DNA", Scientific American  
25 Books, New York; Birren et al. (eds) "Genome Analysis: A Laboratory Manual Series", Vols. 1-4, Cold Spring Harbor Laboratory Press, New York (1998); methodologies as set forth in U.S. Pat. Nos. 4,666,828; 4,683,202; 4,801,531; 5,192,659 and 5,272,057; "Cell Biology: A Laboratory Handbook", Volumes I-III Cellis, J. E., ed. (1994); "Current Protocols in Immunology" Volumes I-III  
30 Coligan J. E., ed. (1994); Stites et al. (eds), "Basic and Clinical Immunology"

(8th Edition), Appleton & Lange, Norwalk, CT (1994); Mishell and Shiigi (eds), "Selected Methods in Cellular Immunology", W. H. Freeman and Co., New York (1980); available immunoassays are extensively described in the patent and scientific literature, see, for example, U.S. Pat. Nos. 3,791,932; 3,839,153; 3,850,752; 3,850,578; 3,853,987; 3,867,517; 3,879,262; 3,901,654; 3,935,074; 3,984,533; 3,996,345; 4,034,074; 4,098,876; 4,879,219; 5,011,771 and 5,281,521; "Oligonucleotide Synthesis" Gait, M. J., ed. (1984); "Nucleic Acid Hybridization" Hames, B. D., and Higgins S. J., eds. (1985); "Transcription and Translation" Hames, B. D., and Higgins S. J., eds. (1984); "Animal Cell Culture" Freshney, R. I., ed. (1986); "Immobilized Cells and Enzymes" IRL Press, (1986); "A Practical Guide to Molecular Cloning" Perbal, B., (1984) and "Methods in Enzymology" Vol. 1-317, Academic Press; "PCR Protocols: A Guide To Methods And Applications", Academic Press, San Diego, CA (1990); Marshak et al., "Strategies for Protein Purification and Characterization - A Laboratory Course Manual" CSHL Press (1996); all of which are incorporated by reference as if fully set forth herein. Other general references are provided throughout this document. The procedures therein are believed to be well known in the art and are provided for the convenience of the reader. All the information contained therein is incorporated herein by reference.

### EXAMPLE 1

#### *Identification of alternatively spliced expressed sequences* Background

The etiology of many kinds of cancers, especially those involving multiple genes or sporadic mutations, is yet to be elucidated. Accumulative EST information coming from heterogeneous tissues and cell-types, can be used as a considerable source to understanding some of the events inherent to carcinogenesis.

Although a large number of current bioinformatics tools are used to predict tissue specific genes in general and cancer specific genes in particular,

all fail to consider alternatively spliced variants [Boguski and Schuler (1995) Nat. Genet. 10:369-71, Audic and Claverie (1997) Genome Res. 7:986-995; Huminiecki and Bicknell (2000) Genome Res. 10:1796-1806; Kawamoto et al. (2000) Genome Res. 10:1817-1827]. Alternative splicing is also overlooked by wet laboratory methods such as SAGE and microarray experiments which have been widely used to study gene expression, however remain to be linked to alternative splicing modeling [see Background section and Valculescu et al. (1995) Science 270:484-487; Caron et al. (2001) Science 291:1289-1292 and Schena et al. (1995) Science 270:467-470].

A computational-based approach was developed to identify alternatively spliced transcripts, which are expressed in a temporal and/or spatial pattern. Examples 1-4 below describe the identification of cancer specific alternatively spliced isoforms, which were identified according to the teachings of the present invention.

#### *Experimental procedures and reagents*

**DATA and LEADS alternative splicing modeling** - GenBank version 125 with genomic build # 25 from National Center for Biotechnology Information (NCBI) was used as an input to the LEADS platform as described [Shoshan et al. (2001) Proc. SPIE Microarrays: Optical Technologies and Informatics 4266:86-95; Matloubian (2000) Nat. Immunol. 1:298-304; David et al. (2002) J. Biol. Chem. 277:18084-18090; Sorek et al. (2002) Genome Res. 12:1060-7]. UniGene Build #146 and libraryQuest.txt were obtained from NCBI and Cancer Genome Anatomy Project (CGAP) in National Cancer Institute (NCI), respectively.

**EST tissue information** - EST information was available in web form from Library Browser or Library Finder in NCBI or in the flat file libraryQuest.txt. The file listed 53 tissue sources, 5 histological states (cancer, multiple histology, normal, pre-cancer, and uncharacterized histology), 6 types of tissue preparations (bulk, cell line, flow-sorted, microdissected, multiple preparation, and uncharacterized), and brief descriptions on each library. 5318

libraries were from bulk tissue preparation {including 5000 ORESTES libraries [Camargo et al. (2001) Proc. Natl. Acad. Sci. USA 98:12103-12108]}, 329 from cell lines, 37 flow-sorted, 66 microdissected, 5 multiple preparation, and 1121 were from uncharacterized preparations. Excluding ORESTES libraries, 507 libraries were designated as 'non-normalized' and 100 were designated 'normalized' or 'subtracted' indicating the pretreatment of mRNA before cDNA library construction. A small number of libraries were derived from the same original sample. These were not considered separately. Library counts of ESTs rather than direct EST counts were used to provide semi-quantitative measurements of expression level, since EST counts in some cases reflect the prevalence of ESTs in one or a few particular libraries, and library counts provide better indications across different tissue types when both normalized and non-normalized libraries were analyzed. Such tissue information analyses are limited to those tissues with a sufficient number of libraries. The inclusion of normalized cDNA libraries allowed the examination of genes expressed at low levels.

The ESTs from 'pooled tissue' or 'uncharacterized tissue' were considered as non-conforming in order to maintain the robustness of the results. In addition, 139,243 ESTs that had no library information were considered non-conforming in investigating tissue- or cancer-specific alternative splicing events.

**Results** - Human EST and mRNA sequences aligned against genomic sequences and clustered through Compugen's LEADS platform were used to identify intron boundaries and alternative splicing sites [Shoshan et al. (2001) Proc. SPIE Microarrays: Optical Technologies and Informatics 4266:86-95; Matloubian (2000) Nat. Immunol. 1:298-304; David et al. (2002) J. Biol. Chem. 277:18084-18090; Sorek et al. (2002) Genome Res. 12:1060-7].

20,301 clusters with 2.0 million ESTs contained at least one mRNA sequence, in general agreement with UniGene build #148 with 20,876 such clusters. The remaining EST sequences, which were clustered to unknown

regions of known genes or to unknown genes were not analyzed. Table 1 below provides some statistics about EST and mRNA clustering. 125,115 introns, and 213,483 exons were aligned either with an mRNA, or with ESTs from at least two libraries if there was no RNA aligned to the gene segment.

5 This was effected to exclude possible genomic contamination in expressed sequences, or other EST technology associated faults.

*Table 1*

| EST          | Cluster | RNA          | Cluster |
|--------------|---------|--------------|---------|
| 1            | 963     | 1            | 6527    |
| 2 – 3        | 1457    | 2-3          | 6372    |
| 4 – 7        | 1532    | 4-7          | 6204    |
| 8 – 15       | 1655    | 8-15         | 1915    |
| 16 – 31      | 1879    | 16-31        | 226     |
| 32 – 63      | 2500    | 32-63        | 40      |
| 64 – 127     | 3481    | 64 and above | 17      |
| 128 – 255    | 3240    | Total        | 20301   |
| 256 – 511    | 1406    |              |         |
| 512 – 1023   | 422     |              |         |
| 1024 – above | 1766    |              |         |
| Total        | 20301   |              |         |

10

**EXAMPLE 2***Cluster distribution of alternatively spliced donor and acceptor sites*

Alternative splice events include exon skipping, alternative 5' or 3' splicing, and intron retention, which can be described by the following simplification: a single exon connects to at least two other exons in either the 3' end (donor site) or the 5' end (acceptor site), as shown in Figure 3. Table 2 below lists some statistics of alternative splicing events based on this simplification.

20



*Table 2*

| Alternative donor site | Cluster | Alternative acceptor site | Cluster |
|------------------------|---------|---------------------------|---------|
| 1                      | 3690    | 1                         | 3751    |
| 2                      | 2269    | 2                         | 2388    |
| 3                      | 1348    | 3                         | 1511    |
| 4                      | 760     | 4                         | 799     |
| 5                      | 435     | 5                         | 508     |
| 6 and above            | 566     | 6 and above               | 710     |
| Total                  | 9068    | Total                     | 9667    |

*Distribution analysis* - As described hereinabove a valid donor-acceptor concatenation must be supported by at least one mRNA or by ESTs from at least two different libraries. 8254 clusters were found to have both alternatively spliced donor and acceptor sites. When the lower bound on the number of EST libraries supporting each donor-acceptor concatenation was increased to three, 13,402 alternatively spliced donor sites were shown to be included in 6892 clusters and 15,015 alternatively spliced acceptor sites were shown to be included in 7570 clusters, while 6111 clusters comprised both alternatively spliced donor and acceptor sites.

**EXAMPLE 3**

*Tissue distribution of ESTs and libraries following LEADS alternative splicing modeling*

Cluster analysis performed to identify alternatively spliced ESTs (see Example 2) was further used for tissue information extraction. Table 3 below lists ten tissue types with the largest numbers of ESTs along with those from pooled or uncharacterized tissues.

*Table 3*

| Tissue   | Number of ESTs |        |        | Number of Libraries |        |       |
|----------|----------------|--------|--------|---------------------|--------|-------|
|          | Normal         | Cancer | Total  | Normal              | Cancer | Total |
| Brain    | 93024          | 87803  | 180827 | 30                  | 25     | 55    |
| Lung     | 35455          | 85596  | 121051 | 92                  | 156    | 248   |
| Placenta | 86571          | 27291  | 113862 | 259                 | 3      | 262   |
| Uterus   | 30052          | 71521  | 101573 | 99                  | 107    | 206   |
| Colon    | 23796          | 74998  | 98794  | 274                 | 445    | 719   |
| Kidney   | 42628          | 46811  | 89439  | 9                   | 54     | 63    |
| Skin     | 32436          | 43085  | 75521  | 8                   | 10     | 18    |

149

|                 |        |       |        |     |     |     |
|-----------------|--------|-------|--------|-----|-----|-----|
| Prostate        | 40312  | 27963 | 68275  | 131 | 135 | 266 |
| Mammary gland   | 26509  | 36638 | 63147  | 305 | 665 | 970 |
| Head and neck   | 12354  | 50167 | 62521  | 62  | 800 | 862 |
| Pooled          | 178618 | 992   | 179610 | 15  | 1   | 16  |
| Uncharacterized | 76193  | 9721  | 85914  | 778 | 106 | 884 |

Evidently, ESTs derived from lung, uterus, colon, kidney, mammary gland, head and neck were obtained mainly from cancerous libraries. The distribution of ESTs in normal and cancer libraries in each case was taken into a consideration and used as a parameter for scoring the differential expression annotation.

#### EXAMPLE 4

##### *Identification of putative cancer specific alternatively spliced transcripts*

Alternative splicing events restricted to cancer tissues were identified by looking for any donor-acceptor concatenations exclusively supported by ESTs from cancer tissues. Table 4 below lists six examples for such. An interesting example is the NONO gene (GenBank Accession No: BC003129), represented by 1496 ESTs. The NONO gene has been previously suggested to code for a possible splicing factor [Dong B, Horowitz DS, Kobayashi R, Krainer AR. Nucleic Acids Res (1993) 25;21(17):4085-92]. Its newly discovered restricted expression to cancer tissues suggests that alternative splicing of multiple genes may be regulated during carcinogenesis.

**Table 4**

| mRNA/<br>EST | Uni<br>Gene<br>ID | Pos.        | Total |   | Type | Specific |   | Non-specific |    |   | Possible<br>function         |
|--------------|-------------------|-------------|-------|---|------|----------|---|--------------|----|---|------------------------------|
|              |                   |             | E     | R |      | E        | R | E            |    | R |                              |
|              |                   |             |       |   |      |          |   | C            | N  |   |                              |
| BC003129     | 172207            | 123,<br>237 | 1496  | 8 | d+   | 15       | 1 | 46           | 20 | 3 | Splicing factor<br>candidate |
| NM_018035    | 279851            | 220,<br>301 | 584   | 2 | d-   | 7        | 0 | 21           | 9  | 2 | No known<br>function         |
| AL519365     | 21938             | 474,<br>513 | 162   | 3 | a-   | 8        | 3 | 6            | 1  | 0 | Oxysterol<br>binding         |

150

|           |        |               |     |   |    |   |   |   |   |   |                                 |
|-----------|--------|---------------|-----|---|----|---|---|---|---|---|---------------------------------|
| BF341144  | 155596 | 507,<br>542   | 148 | 1 | a+ | 6 | 0 | 7 | 4 | 1 | BCL2/adenovirus E1B interacting |
| AB009357  | 7510   | 1372,<br>1452 | 205 | 6 | a+ | 7 | 4 | 2 | 4 | 2 | MAPKKK 7                        |
| NM_002382 | 42712  | 57,<br>84     | 165 | 7 | a- | 8 | 1 | 7 | 3 | 6 | MAX protein                     |

One mRNA/EST containing both splicing junctions identifies the cluster. 'Type' – indicates the type of transcript, which was shown to be cancer specific. The following symbols were used, (d) donor site; (a) acceptor site; (+) proximal exon; (-) distal exon. 'Total' – indicates the number of ESTs or mRNAs which were used for analysis. 'Specific/non-specific' – indicates total library number which was used for analysis. All mRNA sequences under 'specific' were from cancer tissues. 'Position' – identifies splicing boundaries on the sequence. E- EST; R-RNA; C- Cancer; N- Normal.

### EXAMPLE 5

#### *Ontological annotation of proteins - data Collection*

##### *Background*

Recent progress in genomic sequencing, computational biology and ontology development has presented an opportunity to investigate broad biological systems

A gene ontology system was developed and specifically used to annotate human proteins. Examples 5-9 below describe the development of an ontology engine, a computational platform for annotation and resultant annotations of human proteins.

Gene Ontology (GO) and gene association files were obtained from the Gene Ontology Consortium <http://www.geneontology.org/>. InterPro scan from <http://www.ebi.ac.uk/interpro/>, and enzyme database from <http://expasy.proteome.org.au/enzyme/>. The following databases and versions were used, GenBank release 122.0, SwissProt release 39.0, Enzyme database Release 26.0, InterPro database as of April 6, 2001, NCBI LocusLink data as of March 6, 2001, MEDLINE databases as of April 6, 2001, and the following files from Gene Ontology Consortium: gene\_association.fb (version 1.26, 2001/02/19), gene\_association.mgi (version 1.19, 2001/03/01), gene\_association.sgd (version 1.251, 2001/03/13), gene\_association.pombase (version 1.2, 2000/07/22), ec2go (version 1.2, 2000/10/23), and swp2go (version 1.4, 2000/11/15). 58118 SWISS-Prot proteins have been assigned with at least one GO node by the following sources: 15534 proteins were assigned

with at least a functional GO node by conversion of EC (enzyme nomenclature) to GO node. MGI has assigned 5984 SwissProt proteins with GO nodes (<http://www.mgi.org>). 31869 SwissProt proteins were assigned a GO node using SwissProt keyword correspondence and 33048 SwissProt proteins were assigned GO node by InterPro scanning (<http://www.ebi.ac.uk/interpro/>). The nonredundant protein database was constructed from GenPep file from NCBI, along with proteins collected from the Saccharomyces genome database (SGD) [Dwight et al. (2002) Nucleic Acids Res. 30:69-72] and the Drosophila genome database (Flybase) [The Flybase consortium 2002 Nucleic Acids Res. 30:106-108], with a total number of 670130.

### EXAMPLE 6

#### *Generation of progressive sequence clusters*

A two-stage strategy was used to build a detailed homology map between all proteins in the comprehensive protein database (Example 5). In a first stage, all protein pairs with an E score lower than 0.01 using Blastp with default parameters were cataloged. Table 5 lists the distribution of Blastp results.

*Table 5*

| E score                 | Percentage |
|-------------------------|------------|
| $10^{-10} - 10^{-2}$    | 17.58      |
| $10^{-20} - 10^{-10}$   | 13.81      |
| $10^{-30} - 10^{-20}$   | 11.02      |
| $10^{-40} - 10^{-30}$   | 12.91      |
| $10^{-50} - 10^{-40}$   | 10.24      |
| $10^{-60} - 10^{-50}$   | 5.81       |
| $10^{-70} - 10^{-60}$   | 3.64       |
| $10^{-80} - 10^{-70}$   | 2.65       |
| $10^{-90} - 10^{-80}$   | 2.86       |
| $10^{-100} - 10^{-90}$  | 2.53       |
| $10^{-110} - 10^{-100}$ | 2.18       |
| $10^{-120} - 10^{-110}$ | 1.58       |
| $10^{-130} - 10^{-120}$ | 1.50       |
| $10^{-140} - 10^{-130}$ | 1.13       |
| $10^{-150} - 10^{-140}$ | 1.01       |
| $10^{-160} - 10^{-150}$ | 1.01       |
| $10^{-170} - 10^{-160}$ | 0.92       |
| $10^{-178} - 10^{-170}$ | 0.90       |
| 0.00                    | 6.72       |

In the second stage, all homologous protein pairs were aligned through Needleman-Wunsch algorithm with a global alignment to obtain the percentage of identical amino acids between the two proteins. . BLOSUM62 was used as the substitution matrix. The percentage of identity was defined as the number of amino acids aligned with nonnegative scores divided by the number of amino acids in both aligned and unaligned length of two proteins in the global alignment. Table 6 shows a percent identity distribution of protein pairs following global alignment. Evidently, the majority of protein pairs (i.e., 68.5 %) exhibited identity levels in the range of 10-50 %.

Table 6

| Identity Level | Percentage |
|----------------|------------|
| 0 - 10%        | 5.67       |
| 10 - 20%       | 24.66      |
| 20 - 30%       | 19.94      |
| 30 - 40%       | 10.94      |
| 40 - 50%       | 7.31       |
| 50 - 60%       | 7.09       |
| 60 - 70%       | 7.24       |
| 70 - 80%       | 6.70       |
| 80 - 90%       | 5.98       |
| 90 - 100%      | 4.47       |

**EXAMPLE 7****Text mining.**

Correlations between presence of specific MeSH terms, or specific English words in available text information and Gene Ontology assignments in the training data were obtained. The correlations were then used to predict Gene Ontology for unassigned genes.

**Method** - Non-characters in titles and abstracts, and in definition line of gene records were eliminated and words were stemmed through the Lingua::stem module from [www.cpan.org](http://www.cpan.org). Due to the standardized and curated nature of MeSH terms, MeSH terms were not parsed or stemmed. The frequency of each word in all the available text information was calculated. Words that occurred at least 5 times over the whole text information space

were retained for further studies. This cutoff threshold was used to eliminate rare words, wrong spellings, and sometimes even the base pair sequence present in either the definition lines or abstracts. In addition, an upper limit of word frequency (common words such as 'and', 'gene', 'protein') and a lower  
5 limit of word frequency were defined through repeated training process and manual review. The words within the upper and the lower limits were considered as predictive. Since the correlation between the GO nodes and specific words is positive by nature, negative sentences with words such as 'not' and its variants, such as 'unlikely' or 'unresponsive' were excluded from  
10 consideration.

Genes with GO annotation from other sources such as GO consortium, InterPro scanning or keyword mappings were used as training data to obtain the correlation between specific words and specific GO nodes.

15 The following formula was used.  $S = \log(P(m,g)/P(m)P(g))$ , wherein S is the LOD score for word m – GO g combination, wherein P(m,g) is the frequency of term m and GO node g co-occurrence among all word and GO combinations, P(m) is the frequency of occurrence of term m among all word occurrences, and P(g) is the frequency of occurrence of GO node g among all  
20 GO occurrences.

In order to predict GO node for any specific gene which is linked to one to a few dozen words, the sums of LOD scores from all these words for each possible GO were calculated and sorted, and used for further GO annotation. Multiple MeSH terms – GO correlations were tested and were found to be no  
25 more informative than the single MeSH term – GO correlation, and therefore they were not used.

**Results** - Table 7 below, lists general statistics of text information from publicly available sequence databases.

*Table 7*

|  | MeSH term | Title  | Abstract | Definition line |
|--|-----------|--------|----------|-----------------|
| Number of proteins   | 110608    | 106190 | 113073   | 516952          |
| Number of articles   | 71703     | 77314  | 82654    | n/a             |
| Number of unique words*                                    | 40011     | 18175  | 26630    | 25915           |
| Average number of words per article or per definition line | 19.05     | 2.70   | 11.65    | 6.56            |

A predictive probabilistic model was then applied to create possible GO annotations based on the associated text information. Definition lines of sequence records, MeSH term annotations, titles and abstracts from sequence related publications were modeled separately.

The frequency of association of a specific term with a specific GO node in the training data was examined. Parameters such as boundaries of the frequency of MeSH terms and other words were optimized through the training process, using self-validation and cross validation methods. LOD (logarithm of odds ) scores, defined as the logarithm of the ratio between the association frequency of any term-GO pair and the calculated frequency of the random combination of this pair, were used to indicate the relatedness of certain terms with certain GO node. These LOD scores were found to be correlative with the accuracy of GO prediction, as shown in Figure 5. Text information from titles of MEDLINE records appeared to have more predictive power, in particular at lower LOD scores, than text information from other categories. This suggests that the title tended to summarize the gist of an article in a straightforward manner. MeSH terms had similar predictive capabilities as the abstracts, possibly because the MeSH terms were derived from the abstracts, and thus had similar information contents.

Based on text information, a significant number of proteins were predicted to be associated with one or more GO nodes. Table 8 below, lists the number of proteins with predicted GO node from four types of text information in the three categories of GO. These predicted GO annotations were incorporated in GO process to increase the accuracy of homology-based GO

annotation and to generate de novo annotations.

*Table 8*

|                    | MeSH term | Title | Abstract | Definition Line | Total  |
|--------------------|-----------|-------|----------|-----------------|--------|
| Cellular Component | 57845     | 52094 | 57597    | 514191          | 521396 |
| Molecular Function | 57845     | 54152 | 57632    | 516319          | 523384 |
| Biological Process | 57845     | 53970 | 57631    | 516402          | 523385 |

5

To further enhance the accuracy and coverage of GO annotation process, a computational platform for predicting cellular localization, ProLoc (Einat Hazkani-Covo, Erez Levanon, Galit Rotman, Dan Graur and Amit Novik, a manuscript submitted for publication), was used to predict the cellular localization of individual proteins based on their inherent features such as specific localization signatures, protein domains, amino acid composition, pI, and protein length. Only protein sequences that begin with methionine underwent ProLoc analysis. Thus, 88997 out of 93110 proteins in SwissProt version 39 were analyzed, and 78111 proteins have one to three GO predictions in cellular component category.

15

### **EXAMPLE 8**

#### ***Gene ontology assignment***

Progressive single-linkage clusters with 1 % resolution were generated to assign GO annotations (i.e., nodes) to proteins (see Example 6). Protein clustering and annotation assignment were effected at each level of homology. The resolution was 1 % for global alignment identity (i.e., clustering was first effected at 98 %, then at 97 % and so forth). The resolution was 10 fold for the E score of a BlastP homology pair. For example, clustering was performed at  $10^{-50}$ , then at  $10^{-40}$  and so forth.

25

To examine clustering efficiency and homology transitivity, all homology pairs clustered with at least 90 % identity were examined. At this level, there were a total of 57,004 clusters containing 263,259 protein members.



Among these clusters, 23,231 clusters contained at least three protein members (see Figures 6a-c). The lowest homology pairs had an identity of 46 % while being clustered at 90 % or higher identity levels.

Clusters containing proteins with preassociated or predicted ontological annotations were analyzed and best annotations for individual proteins in the clusters were selected through an error weight calculation. Table 9 below, provides statistics on the number of input gene ontology annotations and the number of output annotations following processing.

Table 9

|                    | Input                                    |                            |                    | Output |
|--------------------|--|----------------------------|--------------------|--------|
|                    | GO annotation, conversion, mapping, etc. | Consortium Enzyme InterPro | Text mining ProLoc |        |
| Cellular Component | 44702                                    |                            | 522179             | 574607 |
| Molecular Function | 85626                                    |                            | 526083             | 580767 |
| Biological Process | 69726                                    |                            | 525842             | 578636 |

Over 85 % of proteins were annotated with one or more GO nodes in each of three GO categories. Table 10 below, analyses the number of proteins annotated at different homology levels, showing that GO annotations were achieved throughout the homology spectra.

Table 10

|                       | Cellular Component | Molecular Function | Biological Process |
|-----------------------|--------------------|--------------------|--------------------|
| Text                  | 32257              | 34137              | 30149              |
| $10^{-2} - 10^{-10}$  | 87967              | 71717              | 74277              |
| $10^{-10} - 10^{-50}$ | 122992             | 70088              | 79318              |
| $10^{-50} - 0.0$      | 98059              | 55132              | 59051              |
| 35% - 75%             | 111130             | 97209              | 108334             |
| 75% - 90%             | 38509              | 68282              | 67429              |
| 90% - 99%             | 38991              | 98576              | 90352              |
| Input GO              | 44702              | 85626              | 69726              |

**EXAMPLE 9*****Statistical validation of ontological annotations***

Gene ontology annotations, which were assigned according to the teachings of the present invention, were assessed by automatic cross-validation.

- 5 One fifth of input of input GO annotations were withheld during the GO annotation process and the resultant annotations were compared with the withheld GO nodes. For each protein, the GO node with the lowest error score was examined. Table 11 below, lists the coverage and accuracy of such representative test.

10

**Table 11**

|                    | <b>Total</b> | <b>Predicted GO</b> | <b>Accurate GO</b> |
|--------------------|--------------|---------------------|--------------------|
| Cellular Component | 7431         | 7186                | 4642               |
| Molecular Function | 12999        | 12864               | 10138              |
| Biological Process | 10811        | 10690               | 8080               |

- Evidently, sample coverage ranged from 96 % to 99 % and the reproducibility was between 65 % and 80 %. The lower reproducibility of GO annotations in the "cellular component" category, as compared with that in the
- 15 other two GO categories was consistent with the notion that a short amino acid segment such as a signal peptide affects significantly protein localization. The presence or absence of such small amino acid segments could not be completely captured in sequence similarity comparisons. Detailed analysis of the validation of data indicated that the accuracy of the annotations correlated
- 20 with the homology levels (data not shown). Manual validation of assigned annotations was performed on a total of 500 annotations and about 85 % – 93 % of annotations were found to be correct. The higher percentage of accuracy in the manual examination over the automatic cross-validation resulted from the incomplete annotation of input GO.

25

**EXAMPLE 10*****Description of data***

Example 10a-e below describe the data table in "Summary\_table" file, on the attached CD-ROM3. The data table shows a collection of annotations of differentially expressed nucleic acid sequences, which were identified according to the teachings of the present invention.

Each feature in the data table is identified by "#".

Each transcript in the data table is identified by:

(i) A Serial number, e.g. "251470" in Example 10a, "445259"-  
10 "445262" in Example 10b. I

(ii) An internal arbitrary transcript accession number, e.g. "N62228\_4" in Example 10a, "BE674469\_0", "BE674469\_0\_124", "BE674469\_1", "BE674469\_1\_124" in Example 10b.

The first number of the internal transcript accession number is shared by  
15 all transcripts which belong to the same contig, and represent alternatively spliced variants of each other, e.g. "BE674469" in "BE674469\_0", "BE674469\_0\_124", "BE674469\_1", "BE674469\_1\_124" in Example 10b.

The second number of the internal transcript accession number is an internal serial transcript number of a specific contig, e.g. "\_0" or "\_1" in  
20 "BE674469\_0", "BE674469\_0\_124", "BE674469\_1", "BE674469\_1\_124" in Example 10b.

The third number of the internal transcript accession number is optional, and represents the GenBank database version used for clustering, assembly and annotation processes. Unless otherwise mentioned, GenBank database  
25 version 126 was used. "124" indicates the use of GenBank version 124, as in "BE674469\_1\_124" of Example 10b.

"ProDG" following the internal accession number indicates an EST sequence data from a proprietary source, e.g., Examples 3d and 3e.

"han" represents the use of GenBank version 125. This version was  
30 used in the annotation of lung and colon cancer specific expressed sequences.

“lab” indicates expressed sequences which differential pattern of expression has been confirmed in the laboratory.

Transcript accession number identifies each sequence in the nucleotide sequence data files “Transcripts\_nucleotide\_seqs\_part1”,  
5 “Transcripts\_nucleotide\_seqs\_part2”, “Transcripts\_nucleotide\_seqs\_part3” and “Transcripts\_nucleotide\_seqs\_part4” on CD-ROMs 1 and 2, and in the respective amino acid sequences data file “Protein.seqs” on CD-ROM2. Of note, some nucleotide sequence data files of the above, do not have respective amino acid sequences in the amino acid sequence file “Protein.seqs” attached  
10 on CD-ROM2.

Additional lines of the file contain the following information:

“\*” indicates optional fields; “\*\*” indicates repeatable features.

“#EST” represents a list of GenBank accession numbers of all expressed sequences (ESTs and RNAs) clustered to a contig, from which a  
15 respective transcript is derived. The GenBank accession numbers of these expressed sequences are listed only for the first transcript in the contig, e.g. “#EST BC006216,BE674469,BE798748,NM032716” in Example 10b. The rest of the transcripts derived from the same contig, are indicated by an #EST field marked with “the same”.

20 Expressed sequences, marked with “ProDGyXXX”, e.g., “ProDGy933” in Example 10d, and expressed sequences, marked with “GeneID XXX”, e.g., “GeneID1007Forward” in Example 10e, are proprietary sequences which do not appear in GenBank database. These sequences are deposited in the nucleotide sequence file “ProDG\_seqs” in the attached CD-ROM2.

25 Data pertaining to differentially expressed alternatively spliced sequences is presented in the following format:

\*, \*\* “#TAA\_CD” represents the coordinates of the differentially expressed sequence segment. A single number represents a differentially expressed edge, corresponding to the specific junction between 2 exons.  
30 “TAA\_CD” represented by a pair of numbers represents the start and end

positions of a differentially expressed sequence node. For example, "*#TAA\_CD 269 296*" in Example 10a indicates that the transcript identified as N62228\_4 contains a differentially expressed segment, located between the nucleotides at positions 269 and 296.

5       \*, \*\* "*#TAA\_TIS*" contains information pertaining to specific tissue(s), in which the respective transcript is predicted to be expressed differentially. Tumor tissues are indicated accordingly. For example, "*#TAA\_TIS lung Tumor*" indicates that transcript BE674469\_0 in Example 10b is predicted to be differentially expressed in lung tumor tissues.

10       \*, \*\* "*#DN*" represents information pertaining transcripts, which contain altered functional domains, predicted to act in a dominant negative manner. This field lists the description of the functional domain(s), which is altered in the respective splice variants e.g., "*#DN EGF-like domain*" in Example 10a.

15       Functional annotations of transcripts based on Gene Ontology (GO) are indicated by the following format.

      \*, \*\* "*#GO\_P*", annotations related to Biological Process,  
      \*, \*\* "*#GO\_F*", annotations related to Molecular Function, and  
      \*, \*\* "*#GO\_C*", annotations related to Cellular Component.

20       For each category the following features are optionally addressed:

      "*#GOPR*" represents internal arbitrary accession number of the predicted protein corresponding to the functionally annotated transcript. This internal accession number identifies the protein in the amino acid sequence file "Protein.seqs" in the attached CD-ROM2, together with the internal arbitrary  
25       transcript accession number. For example, "*#GOPR human\_281192*" in Example 10a, is a protein sequence encoded by transcript N62228\_4, which appears in the amino acid sequence file "Protein.seqs" in the attached CD-ROM2 and is identified by both numbers, "N62228\_4" and "human\_281192".

      "*#GO\_Acc*" represents the accession number of the assigned GO entry,  
30       corresponding to the following "*#GO\_Desc*" field.

“#GO\_Desc” represents the description of the assigned GO entry, corresponding to the mentioned “#GO\_Acc” field. For example, “#GO\_Acc 7165 #GO\_Desc signal transduction” in Example 10a, means that the respective transcript is assigned to GO entry number 7165, corresponding to signal transduction pathway.

“#CL” represents the confidence level of the GO assignment, when #CL1 is the highest and #CL5 is the lowest possible confidence level.

“#DB” marks the database on which the GO assignment relies on. The “sp”, as in Example 10a, relates to SwissProt Protein knowledgebase, available from <http://www.expasy.ch/sprot/>. “InterPro”, as in Example 10c, refers to the InterPro combined database, available from <http://www.ebi.ac.uk/interpro/>, which contains information regarding protein families, collected from the following databases: SwissProt (<http://www.ebi.ac.uk/swissprot/>), Prosite (<http://www.expasy.ch/prosite/>), Pfam (<http://www.sanger.ac.uk/Software/Pfam/>), Prints (<http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/>), Prodom (<http://prodes.toulouse.inra.fr/prodom/>), Smart (<http://smart.embl-heidelberg.de/>) and Tigrfams (<http://www.tigr.org/TIGRFAMs/>).

“#EN” represents the accession of the entity in the database(#DB), corresponding to the best hit of the predicted protein. For example, “#DB sp #EN NRG2\_HUMAN” in Example 10a means that the GO assignment in this case was based on SwissProt database, while the closest homologue to the assigned protein is depicted in SwissProt entry “NRG2\_HUMAN”, corresponding to protein named “Pro-neuregulin-2” (<http://www.expasy.org/cgi-bin/niceprot.pl?O14511>). “#DB interpro #EN IPR001609” in Example 10c means that GO assignment in this case was based on InterPro database, while the best hit of the assigned protein is to protein family depicted in SwissProt accession number “IPR001609”, corresponding to “Myosin head (motor domain)” protein family (<http://www.ebi.ac.uk/interpro/IEntry?ac=IPR001609>).

The following two fields correspond to the hierarchical assignment of the differentially expressed sequences to a specific tissue(s), based on the EST content and EST libraries' origin within the contig.

\*, \*\* "#SA" indicates that tissue assignment requires a contig,  
5 containing at least 3 ESTs, where at least 80% thereof are assigned to a selected tissue.

\*, \*\* "#RA" indicates that tissue assignment requires a contig derived from at least two different EST libraries, originally constructed from a specific tissue.

10 **Example 10a**

251470 N62228\_4 #EST the\_same #TAA\_CD 269 296 #TAA\_TIS  
ovary , #TAA\_CD 269 296 #TAA\_TIS ovary Tumor, #TAA\_CD 269 296  
#TAA\_TIS skin Tumor, #TAA\_CD 59 269 #TAA\_TIS ovary , #TAA\_CD 59  
269 #TAA\_TIS ovary Tumor, #TAA\_CD 59 269 #TAA\_TIS skin Tumor #DN  
15 EGF-like.domain #GO\_F #GOPR human\_281192 #GO\_Acc 3823 #GO\_Desc  
antibody #CL 2 #DB sp

#EN NRG2\_HUMAN #GO\_P #GOPR human\_281192 #GO\_Acc  
7165 #GO\_Desc signal transduction #CL 2 #DB sp #EN NRG2\_HUMAN

**Example 10b**

20 445259 BE674469\_0 #EST  
BC006216,BE674469,BE798748,NM032716 #TAA\_CD 0 2537 #TAA\_TIS  
lung , #TAA\_CD 0 2537 #TAA\_TIS lung Tumor

445260 BE674469\_0\_124 #124EST  
BC006216,BE674469,BE798748,NM\_032716 #SA Lung Tumor #RA  
25 lung\_cancer

445261 BE674469\_1 #EST the\_same #TAA\_CD 0 2537 #TAA\_TIS  
lung , #TAA\_CD 0 2537 #TAA\_TIS lung Tumor

445262 BE674469\_1\_124 #124EST  
BC006216,BE674469,BE798748,NM\_032716 #SA Lung Tumor #RA  
30 lung\_cancer

**Example 10c**

314251 HUMM7BA\_0 #EST  
 BF804381,BF805793,BF805830,BG978076,HUMM7BA #GO\_C #GOPR  
 human\_313276 #GO\_Acc 16459 #GO\_Desc myosin #CL 2 #DB interpro #EN  
 5 IPR001609 #GO\_F #GOPR human\_313281 #GO\_Acc 3774 #GO\_Desc motor  
 #CL 1 #DB sp #EN Q14786 #GO\_F #GOPR human\_313281 #GO\_Acc 5524  
 #GO\_Desc ATP binding #CL 1 #DB sp #EN Q14786 #GO\_P #GOPR  
 human\_313281 #GO\_Acc 5983 #GO\_Desc starch catabolism #CL 4 #DB sp  
 #EN Q14786 #SA colon, colonic, gut #RA colon\_normal

10 **Example 10d**

723873 AA157684\_T0\_ProDG #EST  
 AA157684,AA157764,AK057980,BF355351,ProDGy933  
 #GO\_C #GO\_Acc 0016021 #GO\_Desc "integral membrane protein"  
 #GO\_F  
 15 #GO\_Acc 0005978 #GO\_Desc "glycogen biosynthesis" #GO\_P  
 #GO\_Acc 0003707 #GO\_Desc "steroid hormone receptor"

**Example 10e**

723928 GeneID1007Forward\_T0\_ProDG #EST  
 AC018755CDS1,AC018755mRNA1,AW403840,AY040820CDS0,BF3  
 20 59557,  
 BF896787,BF898989,BF899932,BF900235,BF905509,BI518761,BI756629,B  
 I822428,BI906477,BI906754,BM550096,BM922784,GeneID1007Forward,Ge  
 neID285Forward,ProDGy1006 #GO\_C #GO\_Acc 0005887 #GO\_Desc  
 "integral plasma membrane protein" #GO\_F #GO\_Acc 0007267 #GO\_Desc  
 25 "cell-cell signaling"#GO\_P #GO\_Acc 0005530 #GO\_Desc "lectin"

**EXAMPLE 11****Description of the sequence files on the enclosed CD-ROM**

The sequences in the CD-ROM sequence files are in FastA text format.  
 30 Each transcript sequence starts with ">" mark, followed by the transcript internal



accession number. The proprietary ProDG EST sequences starts with ">" mark, followed by the internal sequence accession. An example of the sequence file is presented below.

**Example 11a**

5 >R42278\_0 (SEQ ID NO: 41)  
TGTTTTAGAAATCTCATGATTCCCAGGAAAAAAATTTTAAATTGTGA  
TACAGGTTTGACAGCCTTTTAGTCAAATAAGTTAAAACACACACGC  
AAACTCATTTACTCACTTTGCCATTATAATTCAATCACAAAGAAATT  
TTGGCCAGGCGTGGTGGTTACGCCTGTAATCCCAGCACTTTGGGAGG  
10 CCGAGGCAGGTGGATCACGAGGTCAGGGGATCAAGATCATCCTGGC  
TAACATGTGAAACCCCGTCTCTATTAAAAATAAAAAATTAGCCTGGT  
GTGGTGGCGGGTGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCA  
GCAGAATGGCGTGAACCTCAGGAGGCGGAGCTTGCAGTGAGCCGAG  
ATCGCGCCACTGCACTCCAGCCTGGATGACAGAGCGAGACTCCATC  
15 TCAAAAAAAAAA

**Example 11b**

>GeneID3Reverse #TY RNA #DE ProDGy sequence #DT 18-JUN-1000  
#DR 5 #LN 348 (SEQ ID NO: 1)  
GTGGTTATTACAGCATGGTTCCCAGCCTTACAGTGTCTAAGTGCTTC  
20 TCTTGTGTCCTGTAGATGTTGTGAAAAAGAAAAAAACAAAAAATAC  
ACCACACTGTACTTTTTTCCCCCTGCCCCCGTTACTGCCGGTGATTATT  
ATTAAAAATTAGTTTTTTTTCACATCATTATATCTGGCTTCCTATAAAC  
AACAGCCTTAATTCAGTCAAGACTCCCTTTGGGGAATTCATTTTATT  
AAAAATTGGTGTCTGGATACTTCCCTGTACATGCATAAATATGCATG  
25 CATGTACAGAAAGACTGTATGTGTGTGCCTTGACACACACCCATAC  
CTCTCAGAAAAAGTGTTT

**Example 11c**

>ProDGy1339 #OS Homo sapiens #DE ProDGy sequence #DT 26-JUL-  
2002 #TY EST #DR 5 #AC ProDGy1339 #LN 132 (SEQ ID NO: 2)  
30 CAGAAAGCCCAGAGTAGTCCCTGTAAGAAGCTGAGGGGCGCATACC

165

TCTGGGGTTTGGGTTCCTTCAGGGAAGCGAAGGGAGATGACCTCTT  
TCCAGGCTGGGGACCAAGAGGGCTCCCTAGAAGATATTA

### EXAMPLE 12

5

#### *Description of the CD-ROM content*

The CD-ROMs enclosed herewith contain the following files:

##### *CD-ROM1 (two file):*

1. "Transcripts\_nucleotide\_seqs\_part1", containing nucleotide  
sequences of all the transcripts based on genomic production of GenBank  
10 version 126.

2. "GC\_old.txt" includes the title of the invention and reference  
numbers.

##### *CD-ROM2 (4 files):*

1. "Transcripts\_nucleotide\_seqs\_part2", containing nucleotide  
15 sequences of all the transcripts based on expressed production of GenBank  
version 126 (in cases where no genomic data support was available).

2. "Transcripts\_nucleotide\_seqs\_part3.new", containing nucleotide  
sequences of all the transcripts based on GenBank versions 124, 125, and  
transcripts containing ProDG proprietary sequences.

20 3. "Protein.seqs", containing all the amino acid sequences encoded  
by the transcripts of the invention.

4. "ProDG\_seqs", containing the proprietary EST sequences.

##### *CD-ROM3 (1 file):*

1. "Summary\_table.new", containing all the annotation information,  
25 as described in Example 10.

##### *CD-ROM4 (1 file):*

1. "SEQUENCE LISTING.txt" contains all the sequences which are  
presented in the Examples section of the instant application.

**EXAMPLE 13*****In-vitro confirmation of Differentially Expressed Transcripts******Experimental procedures and reagents***

In-vitro confirmation of in-silico obtained differentially expressed polynucleotide sequences was effected utilizing laboratorial methodologies, based on nucleotide hybridization including northern analysis, RT-PCR and real-time PCR.

***RNA preparation*** - Total RNA was isolated from the indicated cell lines or tumor tissues using the Tri-Reagent (Molecular Research Center Inc.) following the manufacturer's recommendations. Poly(A) RNA was purified from total RNA using oligo(dT)<sub>25</sub> Dynabeads (Dyna).

***Northern blotting*** - 20 µg of total RNA or 2µg of poly(A) RNA were electrophoresed on 1% agarose gels containing formaldehyde, and blotted onto Nytran Super Charge membranes (Schleicher & Schuell). Hybridization was carried out using a DNA probe (SEQ ID NO: 3) in EZ-Hybridization Solution (Biological Industries, Beit Haemek, Israel) at 68°C for 18 hrs. The membranes were rinsed twice with 2XSSC, 0.1% SDS at room temperature, followed by two washes with 0.1XSSC, 0.1%SDS at 50°C. Autoradiograms were obtained by exposing the membranes to X-ray films.

**RT-PCR analysis** - Prior to RT reactions, total RNA was digested with DNase (DNA-free<sup>TM</sup>, Ambion) in the presence of RNasin. Reverse transcription was carried out on 2 µg of total RNA, in a 20 µl reaction, using 2.5 units of Superscript II Reverse Transcriptase (Bibco/BRL) in the buffer supplied by the manufacturer, with 10 pmol of oligo(dT)<sub>25</sub> (Promega), and 30 units of RNasin (Promega). RT reactions were standardized by PCR with GAPDH-specific primers, for 20 cycles. The calibrated reverse transcriptase samples were then analyzed with gene-specific primers either at 35 cycles, or at lower cycles (15 and 20 cycles). PCR products of lower number of cycles were visualized by southern blotting, followed by hybridization with the appropriate probe (the same PCR product).

**Real-Time RT-PCR** - Total RNA samples were treated with DnaseI (Ambion) and purified with Rneasy columns (Qiagen). 2 µg of treated RNA samples were added into 20 µl RT-reaction mixture including. RT-PCR end product 200 units SuperscriptII (Invitrogen), 40 units RNasin, and 500pmol oligo dT. All components were incubated for 1 hr at 50°C and then inactivated by incubation for 15 min at 70°C. Amplification products were diluted, 1:20, in water. 5µl of diluted products were used as templates in Real-Time PCR reactions using specific primers and the intercalating dye Sybr Green.

The amplification stage was effected as follows, 95 °C for 15 sec, 64 °C for 7 sec, 78 °C for 5 sec and 72 °C for 14 sec . Detection was effected using Roch light cycler detector. The cycle in which the reactions achieved a threshold level of fluorescence was registered and served to calculate the initial transcript copy number in the RT reaction. The copy number was calculated using a standard curve created using serial dilutions of a purified amplicon product. To minimize inherent differences in the RT reaction, the resulting copy number was normalized to the levels of expression of the housekeeping genes Proteasome 26S subunit (GenBank Accession number D78151) or GADPH (GenBank Accession number: AF261085).

**Semiquantitative PCR** - RT-PCR reaction was performed with sample

specific primers, for 16 cycles. PCR products were used as probes. Labeling procedure was carried out using "Random primer DNA labeling mix" according to manufacturer's instructions (Cat. No: 20-101-25). Briefly, 25ng of template DNA were denatured by heating to 100 °C for 5 minutes, and then chilled on ice for 5 minutes. Labeling solution contained 11  $\mu$ l of denatured DNA, 4  $\mu$ l of labeling mix solution (Biological industries), 5  $\mu$ l of  $^{32}$ (p)dCTP (Amersham, Pharmacia, AA0005). Labeling was effected for 10 minutes in 37°C. Removal of unincorporated nucleotides was effected using Sephadex G-50 columns. Prior to hybridization, labeled DNA was denatured by heating to 100°C for 5 minutes and then rapidly cooled on ice.

***Southern blotting*** - PCR products were separated on 1.5% agarose gel and size separated. The gel was denatured by two consecutive washes for 20 min in 1X denaturation buffer, containing 1.5M NaCl, 0.5M NaOH. Thereafter a neutralization procedure was effected by washing twice for 20 min in 1X neutralization buffer, containing 1.5M NaCl, 0.5m Tris/HCL pH=7.0. Blotting of the denatured DNA to the nylon membrane was performed overnight with 20XSSC. DNA was UV crosslinked (Stratalinker) to a nylon membrane prior to prehybridization step. Prehybridization was performed using EZ-hybridization solution (Biological Industries, Cat no: 01-889-1B) at 68°C for 1 hour. The DNA blot was subjected to Southern hybridization using specific oligonucleotides end-labeled with adenosine 5'-[ $\gamma$ - $^{32}$ P]triphosphate (>5000 Ci/mmol, Amersham Biosciences, Inc.). Hybridization step was effected at 68°C for 16 hours.

Following hybridization the membrane was washed at gradually increasing stringent conditions: twice in 2X SSC, 0.1%SDS, for 15 min. at room temperature and twice in 0.1XSSC, 0.1%SDS, for 15 min, at 60°C. Radioactive signal was visualized by autoradiography.

#### EXAMPLE 14

##### *Colorectal cancer specific expression of AA535072*

AA535072 (SEQ ID NO: 39) is a common sequence feature to a series of overlapping sequences (SEQ ID NOs: 4, 24-28) with predicted amino acid sequences provided in SEQ ID NOs: 35-38.

The indicated tissues and cell lines were examined for AA535072 (SEQ ID NOs: 39) expression by RT-PCR analysis. Primers for AA535072 were GTGACAGCCAGTAGCTGCCATCTC (SEQ ID NO: 5) and TCCGTTTCTAGCGGCCAGACCTTT (SEQ ID NO: 6). PCR reactions were denatured at 94 °C for 2 minutes followed by 35 cycles at 94 °C for 30 sec, 64 °C for 30 sec and 72 °C for 60 sec. All PCR products were separated on an ethidium bromide stained gel.

As shown in Figure 7 amplification yielded a major PCR product of 1000 bp. Evidently, AA535072 expression was limited to colorectal cancer tissues; adenocarcinoma, colon carcinoma cell line and colon carcinoma Duke A cells. Since colon carcinoma Duke A cells represent an early stage of colon cancer progression, differentially expressed AA535072 can be used as a putative marker of polyps and benign stages of colon cancer. Furthermore, corresponding protein products (SEQ ID NOs: 35-38) may be utilized as important colon cancer specific diagnostic and prognostic tools.

20

### EXAMPLE 15

#### *Bone Tumor Ewing's Sarcoma Specific expression of AA513157*

##### *(SEQ ID NO: 7)*

The indicated tissues and cell lines were examined for AA513157 (SEQ ID NO: 7) expression by RT-PCR analysis. Primers for SEQ ID NO: 7 were GAAGGCAGGCGGATGCTACC (SEQ ID NO: 8) and AGCCTTCCACGCTGTACACGCCA (SEQ ID NO: 9). PCR reactions were denatured at 94 °C for 2 minutes followed by 35 cycles at 94 °C for 30 sec, 64 °C for 30 sec and 72 °C for 45 sec. All PCR products were separated on an ethidium bromide stained gel.

As shown in Figure 8, amplification reaction yielded a specific PCR product of 600 bp. As shown in Figure 8, in the presence of reverse transcriptase (indicated by +) high expression of AA513157 was evident in both samples of Ewing sarcoma, while only residual expression of AA513157 was seen in Ln-Cap cells, brain and splenic adenocarcinoma.

To substantiate these, Northern blot analysis of AA513157 was effected. The following primers were used, GAAGGCAGGCTGGATGCTACC (SEQ ID NO: 10), GGTAGTATAACCGGGCTCTGT (SEQ ID NO: 11). Figure 9 illustrates RNA expression of AA513157 in various tissues. Several transcripts were evident upon Northern analysis: two major transcripts of 800 bp and 1800 bp from ployA RNA preparation and total RNA preparation, respectively. Expression of both transcripts was limited to the Ewing sarcoma cell line. Low expression of the 1800 bp transcript was evident in Bone Ewing sarcoma tissue as well.

These results corroborate AA513157 as a putative Ewing sarcoma marker and a putative pharmaceutical target.

### EXAMPLE 16

#### *Colorectal cancer specific expression of AA469088*

AA469088 (SEQ ID NO: 40) is a common sequence feature to a series of overlapping sequences (SEQ ID NOs: 12 and 29-31).

The indicated tissues and cell lines were examined for AA469088 (SEQ ID NO: 40) expression by semi quantitative RT-PCR analysis. Primers for AA469088 were CATATTTCACTCTGTTCTCTCACC (SEQ ID NO: 13) and CAGAATGGGATTATGGTAGTCTATCT (SEQ ID NO: 14). PCR reactions were effected as follows: 14 cycles at 92 °C for 20 sec, 59 °C for 30 sec and 68 °C for 45 sec. The PCR products were size separated on agarose 1.5% gel, and undergone Southern blot analysis using the PCR products as specific probe, as described in details in Example 13. The visualization of the hybridization signal of the PCR products was performed by autoradiogram exposure to X-ray film.

171

As shown in Figure 10 amplification reaction yielded a major PCR product of 484 bp. Evidently, AA469088 expression was limited to colorectal tumor tissues, normal colon and adenocarcinoma with only minor expression in the spleen and kidney.

5

### **EXAMPLE 17**

#### ***HUMMCDR - A lung cancer specific marker***

Real-time quantitative RT-PCR was used to measure the mRNA steady state levels of HUMMCDR (SEQ ID NO: 15). The following primers were used CTTCAATTGGATTATGTTGACCTCTAC (SEQ ID NO: 16) and CACTATAGGCAACCAGAACAATGTC (SEQ ID NO: 17).

10

Real-time PCR analysis (Figure 11) indicates that SEQ ID NO: 15 is specifically expressed in lung squamous cell carcinoma with an evident 2-10 fold higher expression than in normal lung samples.

15

### **EXAMPLE 18**

#### ***SEQ ID NO: 18 - A lung cancer specific transcript***

Real-time quantitative RT-PCR was used to measure the mRNA steady state levels of SEQ ID NO: 18. The following primers were used GCGAGGACCGGGTATAAGAAGC (SEQ ID NO: 19) and TCGGCTCAGCCAAACACTGTCAG (SEQ ID NO: 20).

20

Real-time PCR analysis indicates that SEQ ID NO: 18 is specifically expressed in lung adenocarcinoma samples and in lung alveolus cell carcinoma (Figure 13).

25

### **EXAMPLE 19**

#### ***SEQ ID NO: 21 - A lung cancer specific transcript***



172

Real-time quantitative RT-PCR was used to measure the mRNA steady state levels of SEQ ID NO: 21. The following primers were used GCTTCGACCGGCTTAGAACT (SEQ ID NO: 22) and GGTGAGCACGATACGGGC (SEQ ID NO: 23).

5 Real-time PCR analysis indicates that SEQ ID NO: 21 is specifically expressed in small lung cell carcinoma and in adenocarcinoma (Figure 14).

### EXAMPLE 20

#### *HSGPGI – A lung cancer specific transcript*

10 Real-time quantitative RT-PCR was used to measure the mRNA steady state levels of HSGPGI (SEQ ID NO: 32). The following primers were used GAGCCCTGTGCGCCGCTCAGATGTG (SEQ ID NO: 33) and AGCCCAAGTTGAATCACCAACCAG (SEQ ID NO: 34).

15 As shown in Figure 12, real-time PCR analysis exhibited specific expression of SEQ ID NO: 32 in lung adenocarcinoma and lung squamos cell carcinoma, as compared to the expression in normal lung tissue (2-25 fold).

### EXAMPLE 21

#### *Comparative analysis of human and mouse alternatively spliced exons*

##### *Rationale and Experimental Procedures*

Alternatively spliced internal exons were identified as described hereinabove [Sorek (2002) Genome Res. 12:1060-1067], essentially screening for reliable exons according to canonical splice sites and discarding possible  
25 genomic contamination events. A constitutively spliced internal exon was defined as an internal exon when supported by at least 4 sequences, for which no alternative splicing was observed. Alternatively, a spliced internal exon was defined as such if there was at least one sequence that contained both the internal exon and the 2 flanking exons (exon inclusion), and at least one  
30 sequence which contained the two flanking exons without the middle one (exon

skipping).

To identify exons which are conserved in mice, mouse ESTs (from GenBank version 131) were aligned to the human genome using a spliced alignment model which allows opening of long gaps. Single hits of mouse  
5 expressed sequences to the human genome shorter than 20 bases, or having less than 75% identity to the human genome, were discarded.

To determine if the borders of a human intron, which define the borders of the flanking exons, were conserved in mice, a mouse EST spanning the same intron-borders, while aligned to the human genome, was sought. Only mouse  
10 EST sequences which exhibited alignment of at least 25 bp on each side of the exon-exon junction were used. In addition, this mouse EST was sought to span an intron (i.e., open a long gap) at the same position along the EST, when aligned to the mouse genome.

A human exon-skipping was considered "conserved" in mice if both  
15 splice variants i.e., the variant that skips the exon and the variant that contains the exon, were supported by mouse ESTs.

As shown in Figures 15a-b, in 149 exon-skipping events, both variants were found in mouse ESTs. It will be appreciated though that when the variant which contains the alternatively spliced exon is a rare variant, or a variant  
20 unique to a tissue which is not represented in mouse EST libraries, there may be no mouse EST, which supports such an event. Nevertheless, if the human exon were really conserved in the mouse transcriptome, it would be expected that the DNA genomic sequence would be conserved between the species.

The basic assumption is that although exons are conserved between the  
25 human and mouse genomes to an average level of 85%, introns are conserved to a much lower extend. Therefore, in cases where there was a skipping variant evident in the mouse transcriptome, but there was no mouse EST showing the variant that contains the exon, the sequence of the human exon was aligned to the relevant intron in the mouse genome. The exon was declared conserved  
30 when the following terms were achieved (i) a significant conservation above

80% identity was found, (ii) the alignment spanned the full length of the human exon; and (iii) the exon was flanked by the canonical AG/GT acceptor and donor sites in the mouse genome. Using this approach 94 additional exon-skipping events conserved between human and mouse were identified (see  
 5 Figures 15a-c).

### *Results*

Two exon sequence sets are available (compiled from Gencarta 3.2, gb131):

1. 243 alternatively spliced exons that are conserved between human  
 10 and mouse and are therefore probably functional for the normal organisms' life. The sequences are described by serial number 725901-7261386 in the attached "Summary\_table.new" on CD-ROM3 and listed in the "Transcripts\_nucleotide\_seqs\_part4" file of the attached CD-ROM2.

2. 737 alternatively spliced exons for which no conservation was  
 15 detected. These sequences can represent alternative splicing which is unique to humans and therefore may define human-specific characteristics. Alternatively, there sequences can represent aberrant splicing which causes or is caused by pathological states. The sequences are described by serial number 726387-727860 in the attached "Summary\_table.new" of CD-ROM3 and listed in the  
 20 "Transcripts\_nucleotide\_seqs\_part4" file of the attached CD-ROM2)

For both sequence sets, each alternative splicing is represented by two transcripts, the first represents the variant that skips the alternatively spliced exon and the second represents the variant that contains the exon. Example for the documentation is illustrated hereinunder.

25 #TRS\_SKIP AA325140\_0\_8 #SKIP BM721749 #RETENT  
 AW583868 #MOUSE\_SKIP BC006836 BE226286 BG916003 BI738124  
 BI854105 BI905506

#MOUSE\_RET AA116659 AF026259 AI326485 AW320664  
 BE569976 BG866594 BI412163 BI655027 BI689102 BI689989 BI695087  
 30 BI904805 MUSCAK NM007584

175

#TRS\_RETENT AA325140\_1\_8 #SKIP BM721749 #RETENT  
 AW583868 #MOUSE\_SKIP BC006836 BE226286 BG916003 BI738124  
 BI854105 BI905506

#MOUSE\_RET AA116659 AF026259 AI326485 AW320664  
 5 BE569976 BG866594 BI412163 BI655027 BI689102 BI689989 BI695087  
 BI904805 MUSCAK NM007584

#TRS\_SKIP – indicates if this transcript represents a skipping variant or  
 a retention variant, which includes the exon.

AA325140\_0\_8 – (contig\_name)\_(0 or 1, where 0 is the skipping  
 10 transcript and 1 is the retention one)\_(number of node which represents the  
 exon)

#SKIP – list of human sequences which skip the exon, i.e., match to the  
 “#TRS\_SKIP” transcript.

#RETENT – list of human sequences which contain the exon, i.e., match  
 15 to the “#TRS\_RETENT” transcript.

#MOUSE\_SKIP – list of mouse sequences which skip the exon.

#MOUSE\_RET – list of mouse sequences which contain the exon.

It is appreciated that certain features of the invention, which are, for  
 20 clarity, described in the context of separate embodiments, may also be provided  
 in combination in a single embodiment. Conversely, various features of the  
 invention, which are, for brevity, described in the context of a single  
 embodiment, may also be provided separately or in any suitable  
 subcombination.

25

Although the invention has been described in conjunction with specific  
 embodiments thereof, it is evident that many alternatives, modifications and  
 variations will be apparent to those skilled in the art. Accordingly, it is intended  
 to embrace all such alternatives, modifications and variations that fall within the  
 30 spirit and broad scope of the appended claims. All publications, patents, patent

applications and sequences identified by their accession numbers mentioned in this specification are herein incorporated in their entirety by reference into the specification, to the same extent as if each individual publication, patent, patent application or sequence identified by their accession number was specifically  
5 and individually indicated to be incorporated herein by reference. In addition, citation or identification of any reference in this application shall not be construed as an admission that such reference is available as prior art to the present invention.

***CD-ROM Content***

The following lists the file content of the three CD-ROMs which are enclosed herewith and filed with the application. File information is provided as: File name/bite size/date of creation/operating system/machine format.

## 5        CD-ROM1 (2 file):

1.        "Transcripts\_nucleotide\_seqs\_part1";        594,303,263        bites;  
September, 4, 2002; PC; Internet Explorer

2.        "GC\_old"; 163 bites; January 13, 2004; PC; Notepad

## CD-ROM2 (5 files):

10        1.        "Transcripts\_nucleotide\_seqs\_part2",        132,371,321        bites;  
September, 4, 2002; PC; Internet Explorer

2.        "Transcripts\_nucleotide\_seqs\_part3.new",        27,709,397        bites  
Spetember, 4, 2002; PC; Internet Explorer

3.        "Protein.seqs", 97,838,818 bites; September, 4, 2002; PC;  
15 Internet Explorer

4.        "ProDG\_seqs", 404,161 bites; September, 4, 2002; PC; Internet  
Explorer

5.        "Transcripts\_nucleotide\_seqs\_part4", 5,463,018 bites; April 24,  
2003; PC; Internet Explorer

## 20        CD-ROM3 (1 file):

1.        "Summary\_table", 590,263,484 bites; April 24, 2003; PC;  
Internet Explorer

## CD-ROM4 (1 file):

1.        "SEQUENCE LISTING.txt"; 87,412 bytes; January 18, 2004; PC;  
25 Notepad

## WHAT IS CLAIMED IS:

1. A method of annotating biomolecular sequences according to a hierarchy of interest, the method comprising:

(a) computationally constructing a dendrogram having multiple nodes, said dendrogram representing the hierarchy of interest, wherein each node of said multiple nodes of said dendrogram is annotated by at least one keyword;

(b) computationally assigning each biomolecular sequence of the biomolecular sequences to a specific node of said multiple nodes of said dendrogram to thereby generate assigned biomolecular sequences; and

(c) computationally classifying each of said assigned biomolecular sequences to nodes hierarchically higher than said specific node, thereby annotating biomolecular sequences according to the hierarchy of interest.

2. The method of claim 1, wherein the biomolecular sequences are selected from the group consisting of polypeptide sequences and polynucleotide sequences.

3. The method of claim 2, wherein said polynucleotides are selected from the group consisting of genomic sequences, expressed sequence tags, contigs, complementary DNA (cDNA) sequences, pre-messenger RNA (mRNA) sequences, and mRNA sequences.

4. The method of claim 1, wherein the biomolecular sequences are selected from the group consisting of annotated biomolecular sequences, unannotated biomolecular sequences and partially annotated biomolecular sequences.

5. The method of claim 1, further comprising homology clustering of the biomolecular sequences prior to step (b).
6. The method of claim 1, wherein said dendrogram is selected from the group consisting of a graph, a list, a map and a matrix.
7. The method of claim 1, wherein the hierarchy of interest is selected from the group consisting of a tissue expression hierarchy, a developmental expression hierarchy, a pathological expression hierarchy, a cellular expression hierarchy, an intracellular expression hierarchy, a taxonomical hierarchy and a functional hierarchy.
8. The method of claim 1, wherein each node of said multiple nodes is a parental node in an additional hierarchy of interest.
9. The method of claim 8, further comprising classifying the biomolecular sequences of said parental node according to said additional hierarchy of interest.
10. The method of claim 1, wherein each of the biomolecular sequences is a member of a sequence contig.
11. The method of claim 1, further comprising the step of confirming annotations of said assigned biomolecular sequence in-vivo and/or in-vitro prior to or following step (c).
12. A method of identifying differentially expressed biomolecular sequences, the method comprising:
  - (a) computationally constructing a dendrogram having multiple nodes, said dendrogram representing the hierarchy of interest, wherein each



node of said multiple nodes of said dendrogram is annotated by at least one keyword;

(b) computationally assigning each biomolecular sequence of the biomolecular sequences to a specific node of said multiple nodes of said dendrogram to thereby generate assigned biomolecular sequences;

(c) computationally classifying each of said assigned biomolecular sequences to nodes hierarchically higher than said specific node, to thereby generate annotated biomolecular sequences; and

(d) identifying annotated biomolecular sequences assigned to a portion of said multiple nodes, thereby identifying differentially expressed biomolecular sequences.

13. The method of claim 12, wherein the biomolecular sequences are selected from the group consisting of polypeptide sequences and polynucleotide sequences.

14. The method of claim 13, wherein said polynucleotides are selected from the group consisting of genomic sequences, expressed sequence tags, contigs, complementary DNA (cDNA) sequences, pre-messenger RNA (mRNA) sequences, and mRNA sequences.

15. The method of claim 12, wherein the biomolecular sequences are selected from the group consisting of annotated biomolecular sequences, unannotated biomolecular sequences and partially annotated biomolecular sequences.

16. The method of claim 12, further comprising homology clustering of the biomolecular sequences prior to step (b).

17. The method of claim 12, wherein said dendrogram is selected from the group consisting of a graph, a list, a map and a matrix.

18. The method of claim 12, wherein the hierarchy of interest is selected from the group consisting of a tissue expression hierarchy, a developmental expression hierarchy, a pathological expression hierarchy, a cellular expression hierarchy, an intracellular expression hierarchy, a taxonomical hierarchy and a functional hierarchy.

19. The method of claim 12, wherein each node of said multiple nodes is a parental node in an additional hierarchy of interest.

20. The method of claim 19, further comprising recursively classifying the biomolecular sequences of said parental node according to said additional hierarchy of interest.

21. The method of claim 12, wherein each of the biomolecular sequences is a member of a sequence contig.

22. The method of claim 12, further comprising the step of confirming differential expression of the differentially expressed biomolecular sequences in-vivo and/or in-vitro following step (d).

23. A computer readable storage medium comprising a database stored in a retrievable manner, said database including files each containing data of a specific node of a dendrogram, said data including biomolecular sequence information and biomolecular sequence annotations, wherein said biomolecular sequence annotations are selected from the group consisting of contig description, tissue specific expression, pathological specific expression,

functional features, parameters for ontological annotation assignment, cellular localization, database sequence source and functional alterations.

24. The computer readable storage medium of claim 23, wherein said database further includes information pertaining to generation of said data and/or potential uses of said data.

25. The computer readable storage medium of claim 23, wherein said database includes the files set forth in enclosed CD-ROMs 1, 2 and/or 3.

26. The computer readable storage medium of claim 23, wherein the medium is selected from the group consisting of a magnetic storage medium, an optical storage medium and an optico-magnetic storage medium.

27. The computer readable storage medium of claim 23, wherein said database is a relational database.

28. The computer readable storage medium of claim 23, wherein said database is a hierarchical database.

29. A system for generating a database of annotated biomolecular sequences, the system comprising a processing unit, said processing unit executing a software application configured for:

(a) constructing a dendrogram having multiple nodes, said dendrogram representing a hierarchy of interest, wherein each node of said multiple nodes of said dendrogram is annotated by at least one keyword;

(b) assigning each biomolecular sequence of the biomolecular sequences to a specific node of said multiple nodes of said dendrogram to thereby generate assigned biomolecular sequences;

(c) classifying each of said assigned biomolecular sequences to nodes hierarchically higher than said specific node, to thereby generate annotated biomolecular sequences; and

(d) storing sequence annotations and sequence information of the annotated biomolecular sequences, thereby generating the database of annotated biomolecular sequences.

30. The system of claim 29, wherein the biomolecular sequences are selected from the group consisting of polypeptide sequences and polynucleotide sequences.

31. The system of claim 30, wherein said polynucleotides are selected from the group consisting of genomic sequences, expressed sequence tags, contigs, complementary DNA (cDNA) sequences, pre-messenger RNA (mRNA) sequences, and mRNA sequences.

32. The system of claim 29, wherein the biomolecular sequences are selected from the group consisting of annotated biomolecular sequences, unannotated biomolecular sequences and partially annotated biomolecular sequences.

33. The system of claim 29, wherein said software application is further configured for homology clustering of the biomolecular sequences prior to step (b).

34. The system of claim 29, wherein said dendrogram is selected from the group consisting of a graph, a list, a map and a matrix.

35. The system of claim 29, wherein said hierarchy of interest is selected from the group consisting of a tissue expression hierarchy, a

developmental expression hierarchy, a pathological expression hierarchy, a cellular expression hierarchy, an intracellular expression hierarchy, a taxonomical hierarchy and a functional hierarchy.

36. The system of claim 29, wherein each node of said multiple nodes is a parental node in an additional hierarchy of interest.

37. The system of claim 36, wherein said software application is further configured for classifying the biomolecular sequences of said parental node according to said additional hierarchy of interest.

38. The system of claim 29, wherein each of the biomolecular sequences is a member of a sequence contig.

39. A method of identifying sequence features unique to differentially expressed mRNA splice variants, the method comprising:

- (a) computationally identifying unique sequence features in each splice variant of an alternatively spliced expressed sequences; and
- (b) identifying differentially expressed splice variants of said alternatively spliced expressed sequences, thereby identifying sequence features unique to differentially expressed mRNA splice variants.

40. The method of claim 39, wherein step (b) is effected by qualifying annotations associated with said alternatively spliced expressed sequences.

41. The method of claim 40, further comprising scoring said annotations associated with said alternatively spliced expressed sequences according to:

- (i) prevalence of said alternatively spliced expressed sequences in normal tissues;

185

- (ii) prevalence of said alternatively spliced expressed sequences in pathological tissues;
- (iii) prevalence of said alternatively spliced expressed sequence in total tissues; and
- (iv) number of tissues and/or tissue types expressing said alternatively spliced expressed sequences;

42. The method of claim 39, wherein step (b) is effected by identifying said unique sequence feature.

43. The method of claim 39, wherein said unique sequence feature is selected from the group consisting of a donor-acceptor concatenation, an alternative exon, an exon and a retained intron.

44. The method of claim 39, wherein said identifying unique sequence features in each splice variant of an alternatively spliced expressed sequence is effected by expressed sequence alignment.

45. A computer readable storage medium comprising data stored in a retrievable manner, said data including sequence information of sequence features unique to differentially expressed mRNA splice variants as set forth in files "Transcripts\_nucleotide\_seqs\_part1", "Transcripts\_nucleotide\_seqs\_part2", "Transcripts\_nucleotide\_seqs\_part3.new", "Transcripts\_nucleotide\_seqs\_part4" and/or "Protein.seqs" of enclosed CD-ROMs 1 and/or 2, and sequence annotations as set forth in annotation categories "#TAA\_CD" and/or "#TAA\_TIS", in the file "Summary\_table.new" of enclosed CD-ROM3.

46. The computer readable storage medium of claim 45, wherein said database further includes information pertaining to generation of said data and potential uses of said data.

47. The computer readable storage medium of claim 45, wherein said medium is selected from the group consisting of a magnetic storage medium, an optical storage medium and an optico-magnetic storage medium.

48. The computer readable storage medium of claim 45, wherein said database further includes information pertaining to gain and/or loss of function of said differentially expressed mRNA splice variants or polypeptides encoded thereby.

49. A system for generating a database of sequence features unique to differentially expressed mRNA splice variants, the system comprising a processing unit, said processing unit executing a software application configured for:

(a) identifying unique sequence features in each splice variant of an alternatively spliced expressed sequences; and

(b) identifying differentially expressed splice variants of said alternatively spliced expressed sequences, thereby identifying sequence features unique to differentially expressed mRNA splice variants.

(c) storing the sequence features unique to the differentially expressed mRNA splice variants, thereby generating the database of sequence features unique to differentially expressed mRNA splice variants.

50. The system of claim 49, wherein step (b) is effected by qualifying annotations associated with said alternatively spliced expressed sequences.

51. The system of claim 50, further configured for scoring said annotations associated with said alternatively spliced expressed sequences according to:

(i) prevalence of said alternatively spliced expressed sequences in normal tissues;

- (ii) prevalence of said alternatively spliced expressed sequences in pathological tissues;
- (iii) prevalence of said alternatively spliced expressed sequence in total tissues; and
- (iv) number of tissues and/or tissue types expressing said alternatively spliced expressed sequences;

52. The method of claim 49, wherein step (b) is effected by identifying said unique sequence feature.

53. The system of claim 49, wherein said unique sequence feature is selected from the group consisting of a donor-acceptor concatenation, an alternative exon, an exon and a retained intron.

54. The system of claim 49, wherein said identifying unique sequence features in each splice variant of an alternatively spliced expressed sequence is effected by expressed sequence alignment.

55. A kit useful for detecting differentially expressed polynucleotide sequences, the kit comprising at least one oligonucleotide being designed and configured to be specifically hybridizable with a polynucleotide sequence selected from the group consisting of sequence files "Transcripts\_nucleotide\_seqs\_part1", "Transcripts\_nucleotide\_seqs\_part2", "Transcripts\_nucleotide\_seqs\_part3.new" and/or "Transcripts\_nucleotide\_seqs\_part4" of enclosed CD-ROMs 1 and/or 2 under moderate to stringent hybridization conditions.

56. The kit of claim 55, wherein said at least one oligonucleotide is labeled.



57. The kit of claim 55, wherein said at least one oligonucleotide is attached to a solid substrate.

58. The kit of claim 57, wherein said solid substrate is configured as a microarray and whereas said at least one oligonucleotide includes a plurality of oligonucleotides each being capable of hybridizing with a specific polynucleotide sequence of the polynucleotide sequences set forth in the files "Transcripts\_nucleotide\_seqs\_part1", "Transcripts\_nucleotide\_seqs\_part2" "Transcripts\_nucleotide\_seqs\_part3.new" and/or "Transcripts\_nucleotide\_seqs\_part4" of enclosed CD-ROMs 1 and/or 2.

59. The kit of claim 58, wherein each of said plurality of oligonucleotides is being attached to said microarray in a regio-specific manner.

60. The kit of claim 55, wherein said at least one oligonucleotide is designed and configured for DNA hybridization.

61. The kit of claim 55, wherein said at least one oligonucleotide is designed and configured for RNA hybridization.

62. A method of annotating biomolecular sequences, the method comprising:

(a) computationally clustering the biomolecular sequences according to a progressive homology range, to thereby generate a plurality of clusters each being of a predetermined homology of said homology range; and

(b) assigning at least one ontology to each cluster of said plurality of clusters, said at least one ontology being:

(i) derived from an annotation preassociated with at least one biomolecular sequence of each cluster; and/or

(ii) generated from analysis of said at least one biomolecular sequence of each cluster thereby annotating biomolecular sequences.

63. The method of claim 62, wherein the biomolecular sequences are selected from the group consisting of polynucleotide sequences and polypeptide sequences.

64. The method of claim 62, wherein said homology range is between 99 % - 35 %.

65. The method of claim 62, wherein said analysis of said at least one biomolecular sequence includes literature text mining.

66. The method of claim 62, wherein said analysis of said at least one biomolecular sequence includes cellular localization prediction.

67. The method of claim 62, wherein said analysis of said at least one biomolecular sequence includes homology analysis.

68. The method of claim 62, wherein said at least one ontology is selected from the group consisting of molecular biology, microbiology, developmental biology, immunology, virology, biochemistry, physiology, pharmacology, medicine, bioinformatics, cell biology, endocrinology, structural biology, mathematics, chemistry, medicine, plant sciences, neurology, genetics, zoology, ecology, genomics, cheminformatics, computer sciences, statistics, physics and artificial intelligence.

69. The method of claim 62, wherein said ontology includes a subontology.

190

70. The method of claim 62, further comprising scoring said at least one ontology assigned to a cluster of said plurality of clusters according to:

- (i) a degree of homology characterizing said cluster; and
- (ii) relevance of annotation to information obtained from literature text mining.

71. The method of claim 62, further comprising generating a sequence profile to each cluster of said plurality of clusters following step (b).

72. A system for generating a database of annotated biomolecular sequences, the system comprising a processing unit, said processing unit executing a software application configured for:

- (a) clustering the biomolecular sequences according to a progressive homology range, to thereby generate a plurality of clusters each being of a predetermined homology of said homology range; and

- (b) assigning at least one ontology to each cluster of said plurality of clusters, said at least one ontology being:

- (i) derived from an annotation preassociated with at least one biomolecular sequence of each cluster; and/or

- (ii) generated from analysis of said at least one biomolecular sequence of each cluster, to thereby annotate the biomolecular sequences; and

- (c) storing sequence annotations and sequence information of the annotated biomolecular sequences, thereby generating said database of annotated biomolecular sequences.

73. The system of claim 72, wherein the biomolecular sequences are selected from the group consisting of polynucleotide sequences and polypeptide sequences.

74. The system of claim 72, wherein said homology range is between 99 % - 35 %.

75. The system of claim 72, wherein said analysis of said at least one biomolecular sequence includes literature text mining.

76. The system of claim 72, wherein said analysis of said at least one biomolecular sequence includes cellular localization prediction.

77. The system of claim 72, wherein said analysis of said at least one biomolecular sequence includes homology analysis.

78. The system of claim 72, wherein said at least one ontology is selected from the group consisting of molecular biology, microbiology, developmental biology, immunology, virology, biochemistry, physiology, pharmacology, medicine, bioinformatics, cell biology, endocrinology, structural biology, mathematics, chemistry, medicine, plant sciences, neurology, genetics, zoology, ecology, genomics, cheminformatics, computer sciences, statistics, physics and artificial intelligence.

79. The system of claim 72, wherein said ontology includes a subontology.

80. The system of claim 72 further comprising scoring said at least one ontology assigned to a cluster of said plurality of clusters according to:

- (i) a degree of homology characterizing said cluster; and
- (ii) relevance of annotation to information obtained from literature text mining.

81. The system of claim 72, further comprising generating a sequence profile to each cluster of said plurality of clusters following step (b).

82. A computer readable storage medium comprising a database stored in a retrievable manner, said database including sequence information as set forth in files "Transcripts\_nucleotide\_seqs\_part1", "Transcripts\_nucleotide\_seqs\_part2", "Transcripts\_nucleotide\_seqs\_part3.new" "Transcripts\_nucleotide\_seqs\_part4", and/or "Protein.seqs" of enclosed CD-ROMs 1 and/or 2, and sequence ontological annotations in #GO\_P, #GO\_F and/or #GO\_C annotation categories in file "Summary\_table.new" of enclosed CD-ROM3.

83. The computer readable storage medium of claim 82, wherein said database further includes information pertaining to generation of said data and potential uses of said data.

84. The computer readable storage medium of claim 82, wherein the medium is selected from the group consisting of a magnetic storage medium, an optical storage medium and an optico-magnetic storage medium.

85. A computer readable storage medium, comprising a database stored in a retrievable manner, said database including biomolecular sequence information as set forth in files "Transcripts\_nucleotide\_seqs\_part1", "Transcripts\_nucleotide\_seqs\_part2", "Transcripts\_nucleotide\_seqs\_part3.new" "Transcripts\_nucleotide\_seqs\_part4", and/or "Protein.seqs" of enclosed CD-ROMs 1 and/or 2, and biomolecular sequence annotations as set forth in file "Summary\_table.new" of enclosed CD-ROM 3.

86. The computer readable storage medium of claim 85, wherein said database further includes information pertaining to generation of said data and potential uses of said data.

87. The computer readable storage medium of claim 85, wherein the medium is selected from the group consisting of a magnetic storage medium, an optical storage medium and an optico-magnetic storage medium.

88. The computer readable storage medium of claim 85, wherein said sequence annotations are selected from the group consisting of contig description, position of unique sequence features, tissue specific expression, pathological specific expression, functional features, parameters for ontological annotation assignment, cellular localization, database sequence source and functional alterations.

89. A method of diagnosing colon cancer in a subject, the method comprising identifying in the subject the presence or absence of a biomolecular sequence selected from the group consisting of SEQ ID NOs: 4, 39, 24-28, 35-38, 12 and 29-31 wherein presence of said biomolecular sequence indicates colon cancer in the subject.

90. A method of diagnosing lung cancer in a subject, the method comprising identifying in the subject the presence or absence of a biomolecular sequence selected from the group consisting of SEQ ID NOs: 15, 18, 21 and 32 wherein presence of said biomolecular sequence indicates lung cancer in the subject.

91. A method of diagnosing Ewing sarcoma in a subject, the method comprising identifying in the subject the presence or absence of a biomolecular sequence as set forth in SEQ ID NO: 7, wherein presence of said biomolecular sequence indicates Ewing sarcoma in the subject.

92. A computer readable storage medium comprising data stored in a retrievable manner, said data including sequence information of differentially

expressed biomolecular sequences as set forth in files "Transcripts\_nucleotide\_seqs\_part1", "Transcripts\_nucleotide\_seqs\_part2" "Transcripts\_nucleotide\_seqs\_part3.new", "Transcripts\_nucleotide\_seqs\_part4" and/or "Protein.seqs" of enclosed CD-ROMs 1 and/or 2, and sequence annotations as set forth in annotation categories "SA" and/or "RA", in the file "Summary\_table.new" of enclosed CD-ROM3.

93. The computer readable storage medium of claim 92, wherein said database further includes information pertaining to generation of said data and potential uses of said data.

94. The computer readable storage medium of claim 92, wherein said medium is selected from the group consisting of a magnetic storage medium, an optical storage medium and an optico-magnetic storage medium.

95. The computer readable storage medium of claim 92, wherein said database further includes information pertaining to gain and/or loss of function of said differentially expressed mRNA splice variants or polypeptides encoded thereby.

96. A computer readable storage medium comprising data stored in a retrievable manner, said data including sequence information of biomolecular sequences exhibiting gain of function or loss of function as set forth in files "Transcripts\_nucleotide\_seqs\_part1", "Transcripts\_nucleotide\_seqs\_part2" "Transcripts\_nucleotide\_seqs\_part3.new", "Transcripts\_nucleotide\_seqs\_part4" and/or "Protein.seqs" of enclosed CD-ROMs 1 and/or 2, and sequence annotations as set forth in annotation category "DN", in the file "Summary\_table.new" of enclosed CD-ROM3.

97. The computer readable storage medium of claim 96, wherein said

database further includes information pertaining to generation of said data and potential uses of said data.

98. The computer readable storage medium of claim 96, wherein said medium is selected from the group consisting of a magnetic storage medium, an optical storage medium and an optico-magnetic storage medium.



1/19

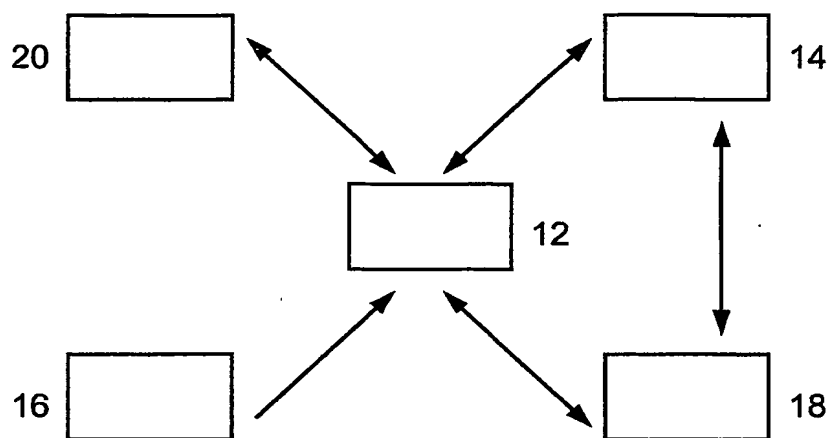


Fig. 1a

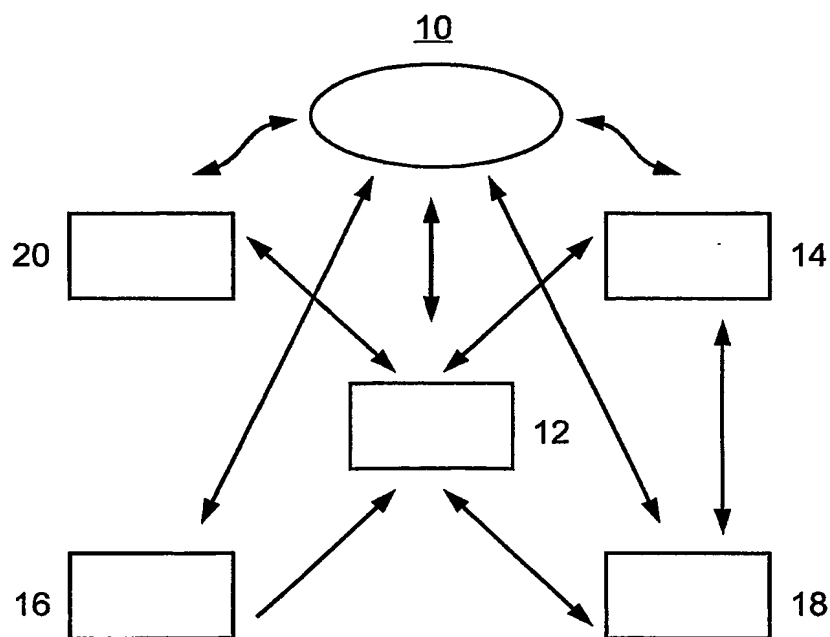


Fig. 1b

2/19

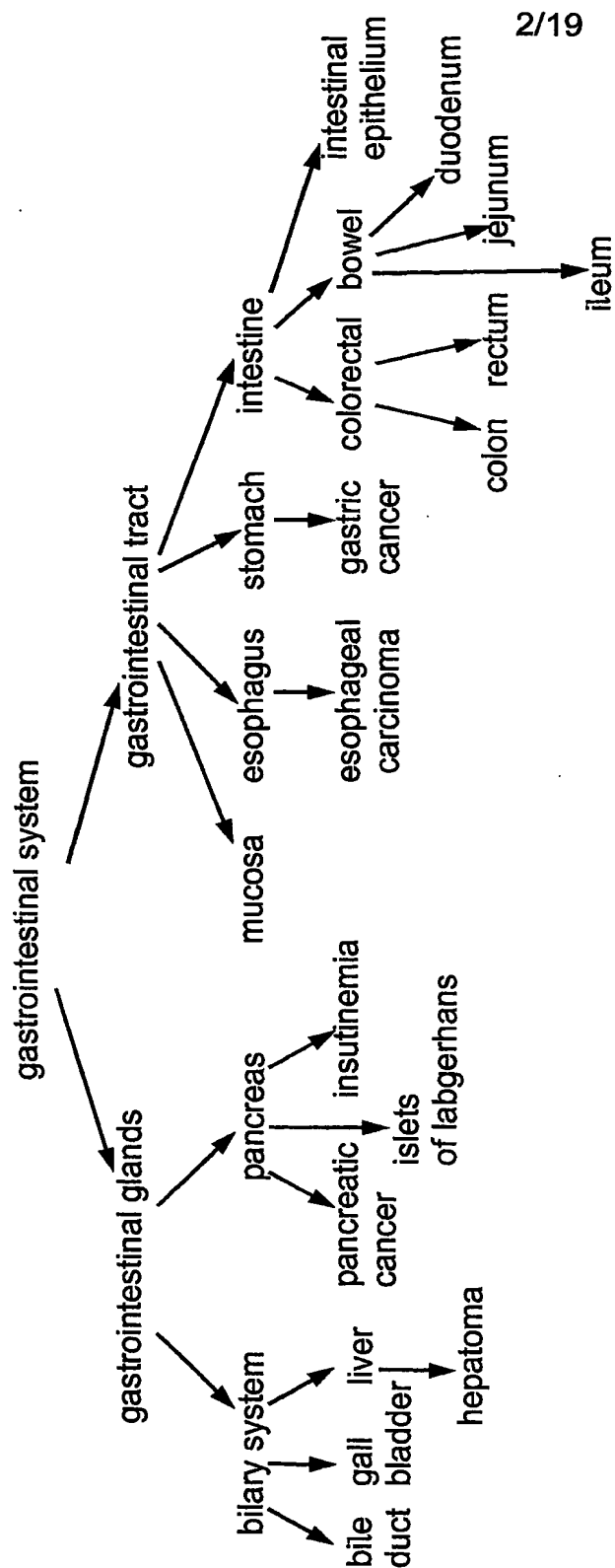


Fig. 2

3/19

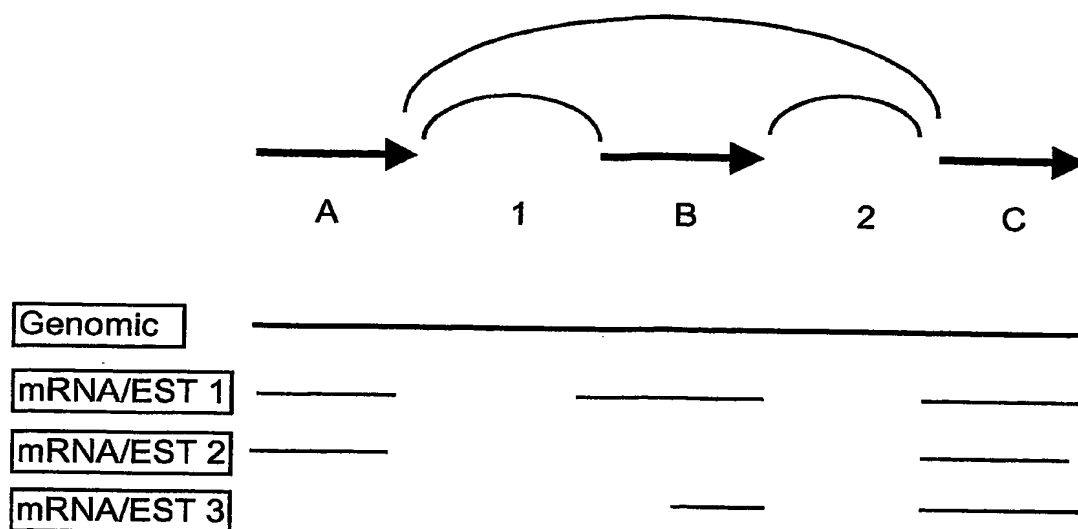


Fig. 3

4/19

- |                                    |                                   |
|------------------------------------|-----------------------------------|
| <b>1 tumor</b>                     | <b>2 endocrine system</b>         |
| <b>1.1 epithelial cell tumors</b>  | <b>2.1 adrenal</b>                |
| <b>1.1.1 carcinoma</b>             | <b>2.1.1 pheocromocytoma</b>      |
| <b>1.1.1.1 adenocarcinoma</b>      | <b>2.2 pancreas</b>               |
| <b>1.1.1.2 lobullar carcinoma</b>  | <b>2.2.1 islets of Langerhans</b> |
| <b>1.2 Mesenchimal cell tumors</b> | <b>2.3 neuroendocrine</b>         |
| <b>1.2.1 sarcoma</b>               | <b>2.3.1 hypothalamus</b>         |
| <b>1.2.1.1 liposarcoma</b>         | <b>2.3.2 carcinoid</b>            |
| <b>1.2.1.2 rhabdomyosarcoma</b>    | <b>2.4 thyroid</b>                |
| <b>1.2.1.3 pnet</b>                |                                   |
| <b>1.2.1.4 ewing sarcoma</b>       | <b>3 vascular tissue</b>          |
| <b>1.3 blood tumors</b>            | <b>3.1 arteries</b>               |
| <b>1.3.1 lymphoma</b>              | <b>3.1.1 aorta</b>                |
| <b>1.3.2 leukemia</b>              | <b>3.2 vein</b>                   |
| <b>1.3.3 myeloma</b>               |                                   |
| <b>1.4 endocrine tumors</b>        |                                   |
| <b>1.4.1 pheocromocytoma</b>       |                                   |
| <b>1.4.2 carcinoid</b>             |                                   |

Fig. 4

5/19

**4 genitourinary system****4.1 urinary system****4.1.1 bladder****4.1.2 kidney****4.2 genital system****4.2.1 women genital system****4.2.1.1 cervix****4.2.1.2 ovary****4.2.1.3 uterus****4.2.1.3.1 endometrium****4.2.2 men gentile system****4.2.2.1 prostate****4.2.2.2 testis****4.2.2.2.1 epididymis****5 muscles****5.1 rhabdomyosarcoma****5.2 tongue****5.3 bladder****5.4 heart****5.5 uterus****6 Blood****6.1 peripheral blood****6.1.1 erythroid line****6.1.2 leukocyte****6.1.2.1 lymphoid system****6.1.2.1.1 lymphoma****6.1.2.1.2 spleen****6.1.2.1.3 thalamus****6.2 stem cells****6.2.1 myeloid****6.2.2 myeloma****6.3 Bone marrow****6.4 leukemia**

Fig. 4 (Cont.)

6/19

**7 nerve system****7.1 CNS, central nervous system****7.1.1 brain****7.1.1.1 cerebrum****7.1.1.2 cerebellum****7.1.1.3 pituitary gland****7.1.1.4 hypothalamus****7.1.1.5 thalamus****7.1.1.6 olfactory****7.1.1.7 Hippocampus****7.1.1.8 amygdala****7.1.1.9 frontal lobe****7.1.1.10 pnet****7.2 Embryonal nerve system****7.2.1 primitive neuroectoderm****7.3 retina****7.3.1 retinoblastoma****8 breast****8.1 ductal breast****8.1.1 ductal carcinoma****8.2 lobular carcinoma****8.3 mammary****9 skeleton****9.1 bone****9.1.1 ewing sarcoma****9.1.2 craniofacial****9.1.2.1 calvarium****9.2 connective tissue****9.2.1 trabeculae****9.2.2 cartilage****10 embryo****10.1 amnion****10.2 chorion****10.3 primitive neuroectoderm****10.4 placenta**

Fig. 4 (Cont.)

|   |                                    |
|---|------------------------------------|
|   | 7/19                               |
| <b>11 exocrine system</b>               | <b>11.4 ear</b>                    |
| <b>11.1 pancreas</b>                    | <b>11.4.1 cochlea</b>              |
| <b>11.1.1 islets of Langerhans</b>      | <b>11.5 eye</b>                    |
| <b>11.2 prostate</b>                    | <b>11.5.1 retina</b>               |
| <b>11.3 salivary gland</b>              | <b>11.5.1.1 retinoblastoma</b>     |
|   | <b>11.5.2 lens</b>                 |
|   | <b>11.6 mouth</b>                  |
|   | <b>11.7 tongue</b>                 |
| <b>13 respiratory system</b>            | <b>12 gastrointestinal system</b>  |
| <b>13.1 nasopharynx</b>                 | <b>12.1 mucosa</b>                 |
| <b>13.2 lung</b>                        | <b>12.2 stomach</b>                |
| <b>13.2.1 small cell lung carcinoma</b> | <b>12.3 intestine</b>              |
|   | <b>12.3.1 colorectal</b>           |
| <b>14 skin</b>                          | <b>12.3.1.1 colon</b>              |
| <b>14.1 dermis</b>                      | <b>12.4 hepatobiliary system</b>   |
| <b>14.1.1 melanocyte</b>                | <b>12.4.1 liver</b>                |
| <b>15 fat tissue</b>                    | <b>12.4.2 biliary system</b>       |
| <b>15.1 liposarcoma</b>                 | <b>12.4.2.1 gall bladder</b>       |
|   | <b>12.5 pancreas</b>               |
|   | <b>12.5.1 islets of Langerhans</b> |

Fig. 4 (Cont.)

8/19

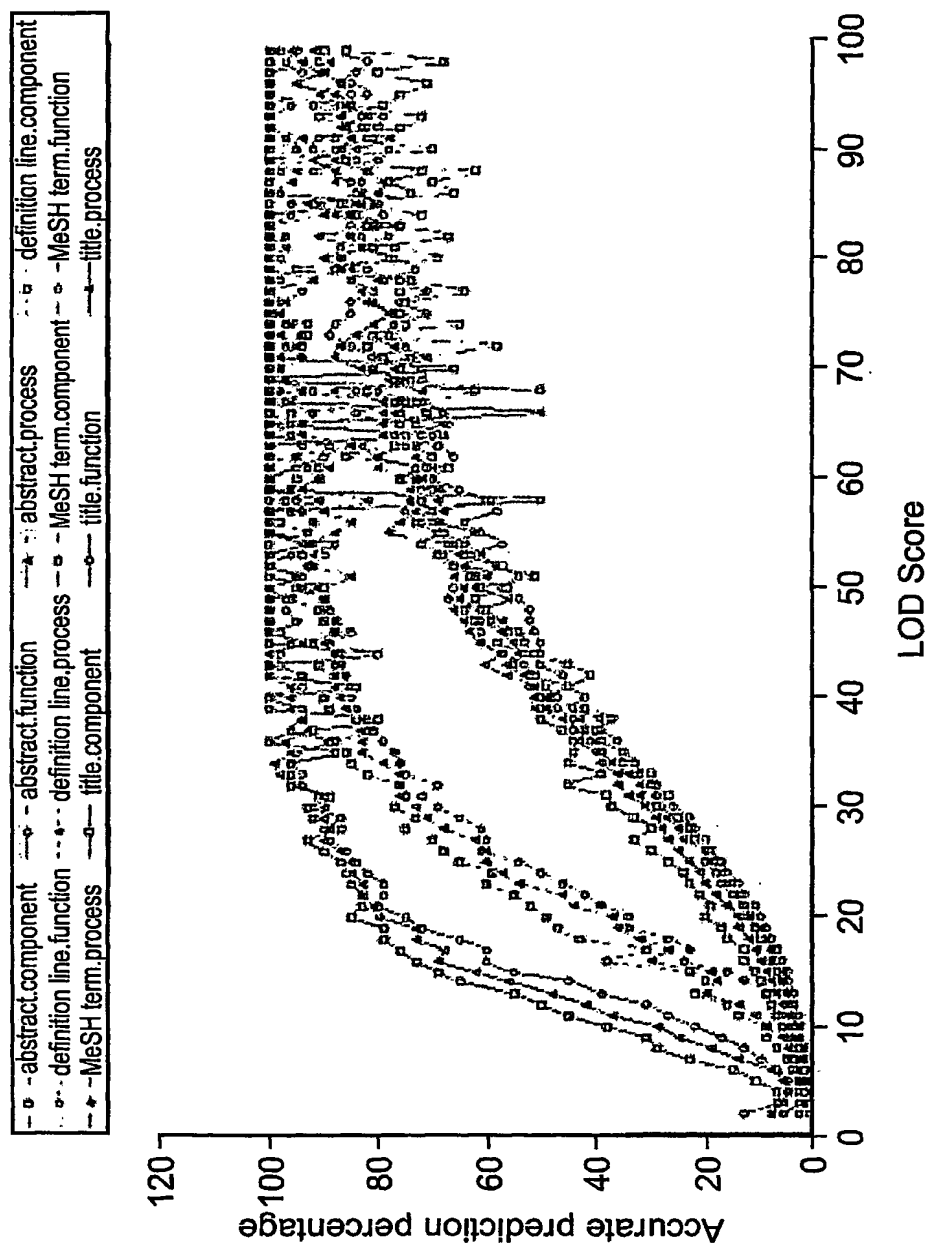


Fig. 5



9/19

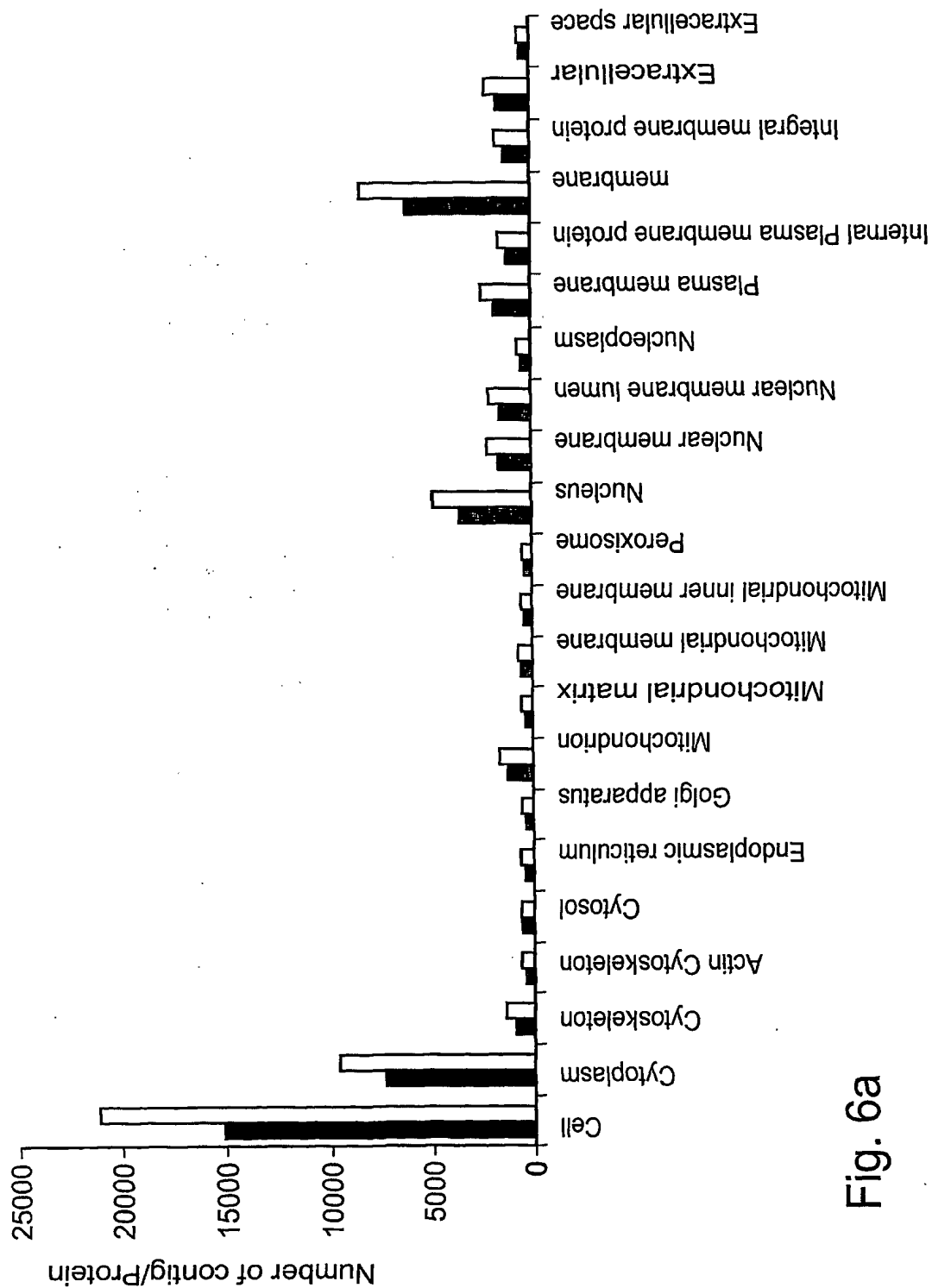


Fig. 6a

10/19

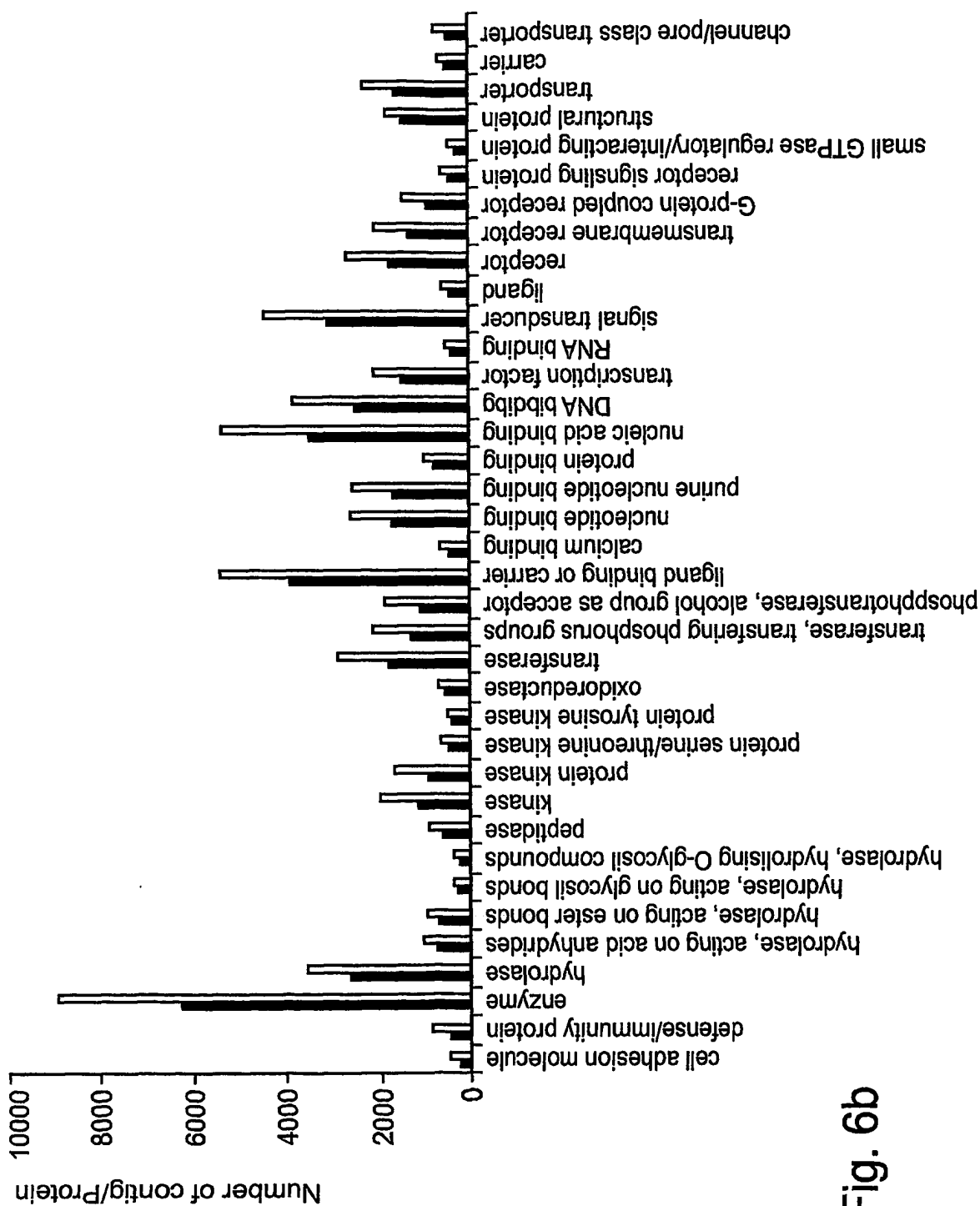


Fig. 6b

11/19

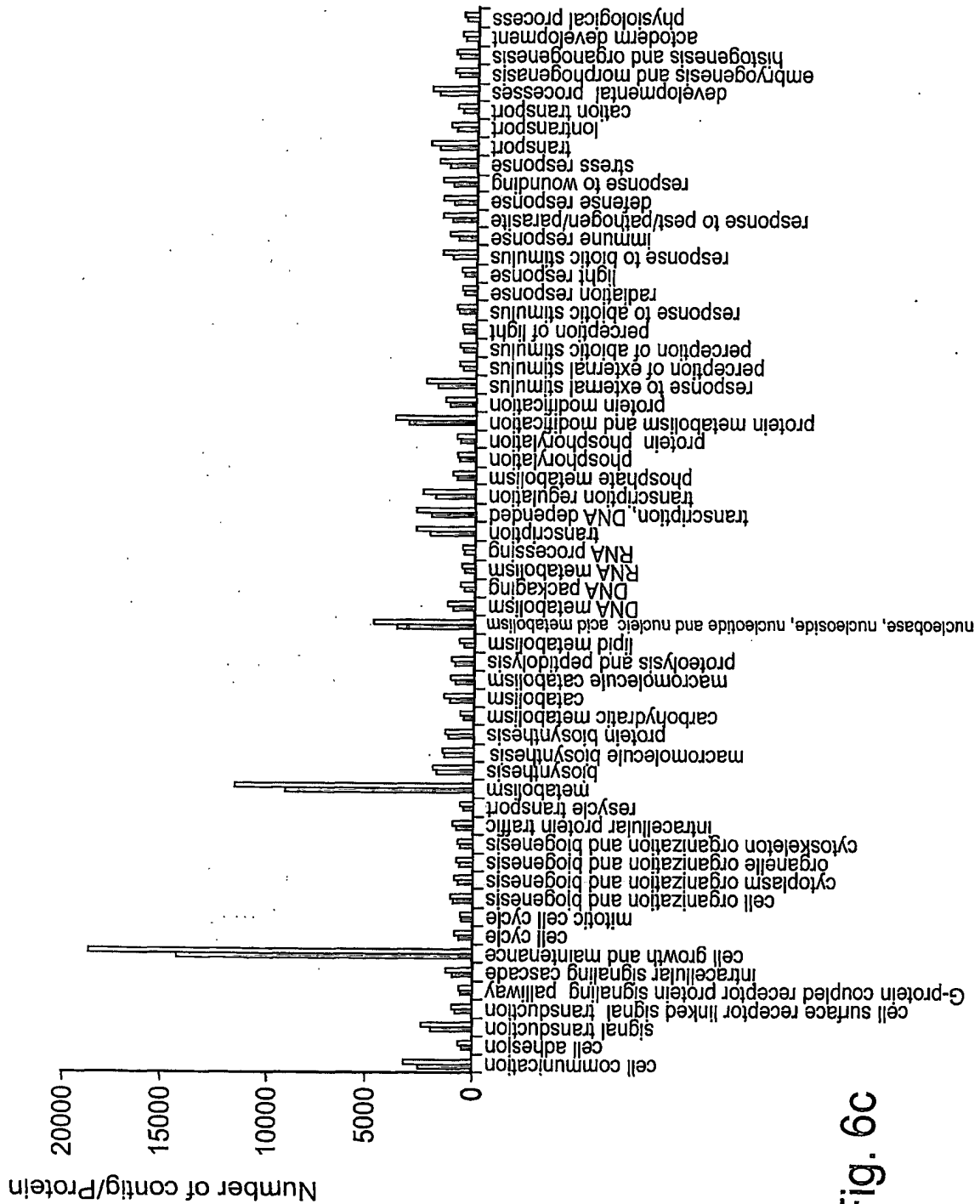


Fig. 6c

12/19

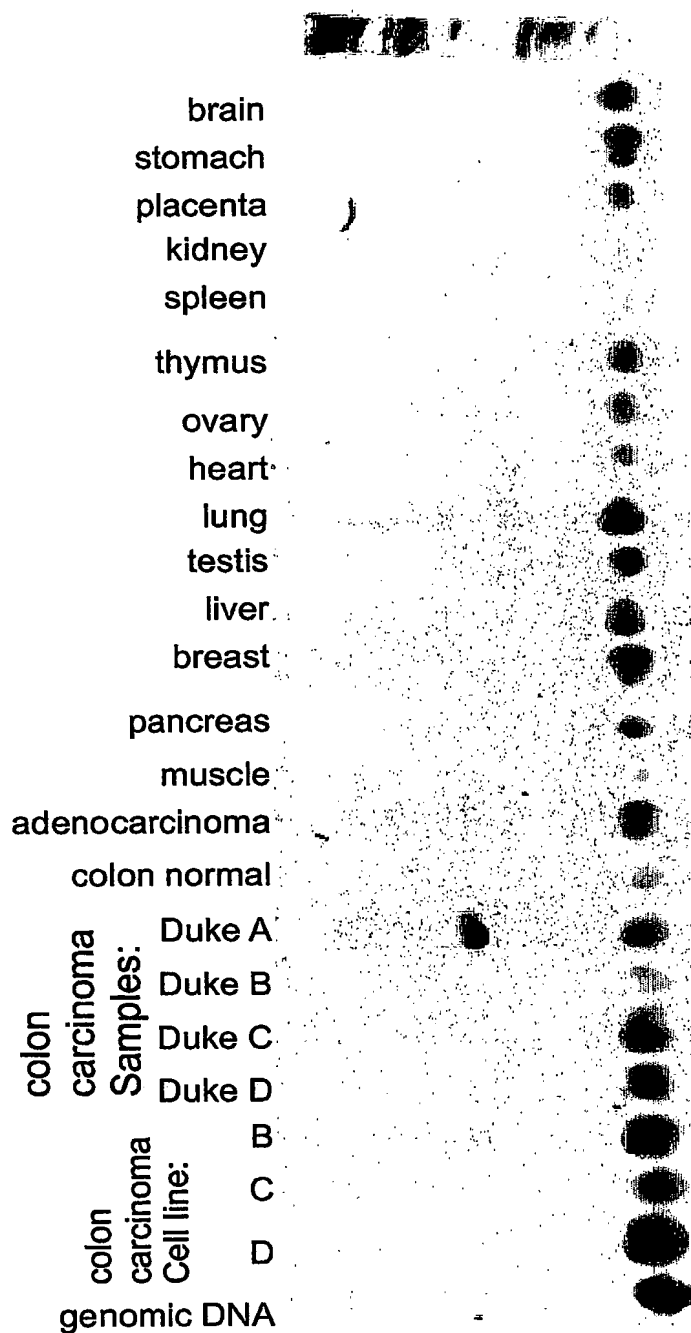


Fig. 7

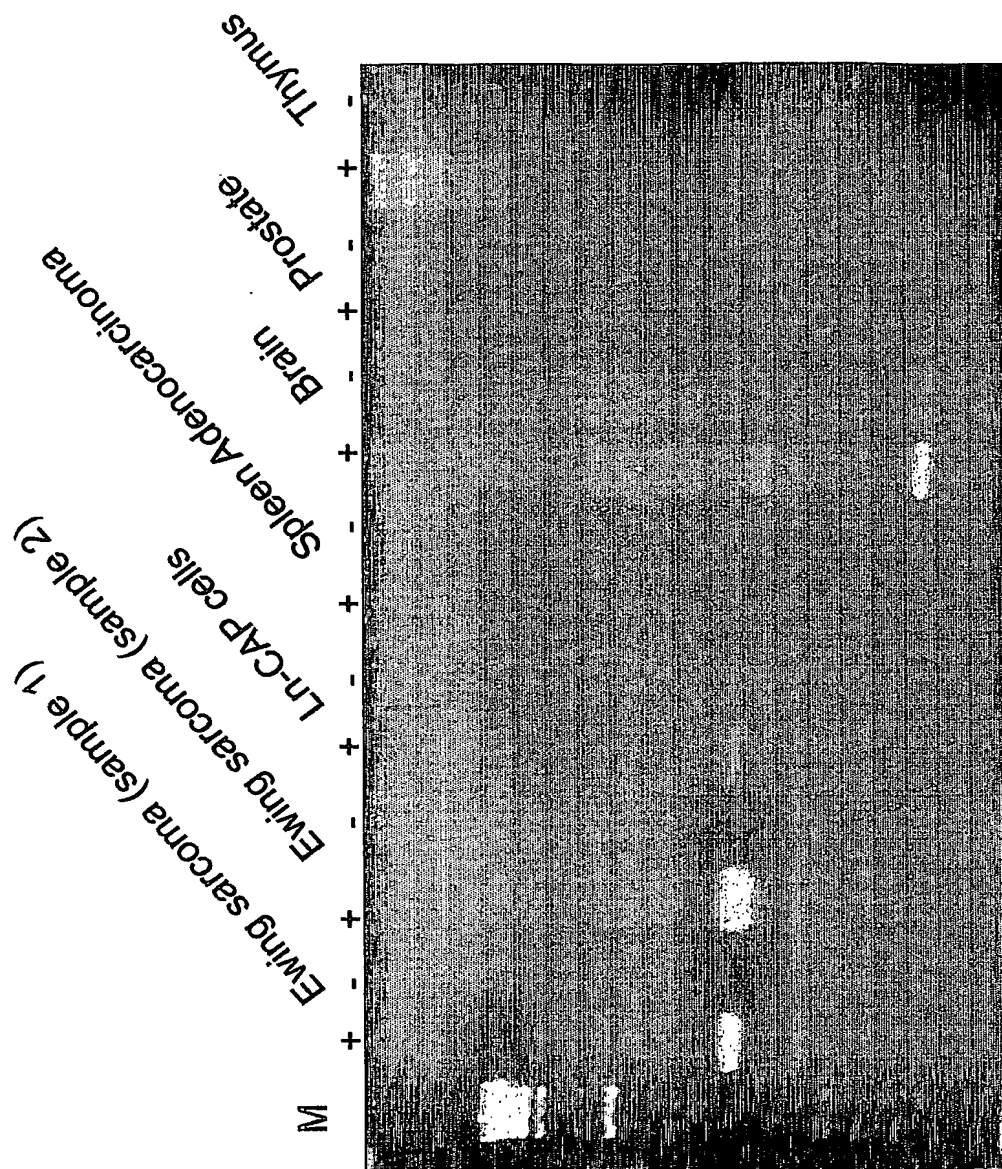
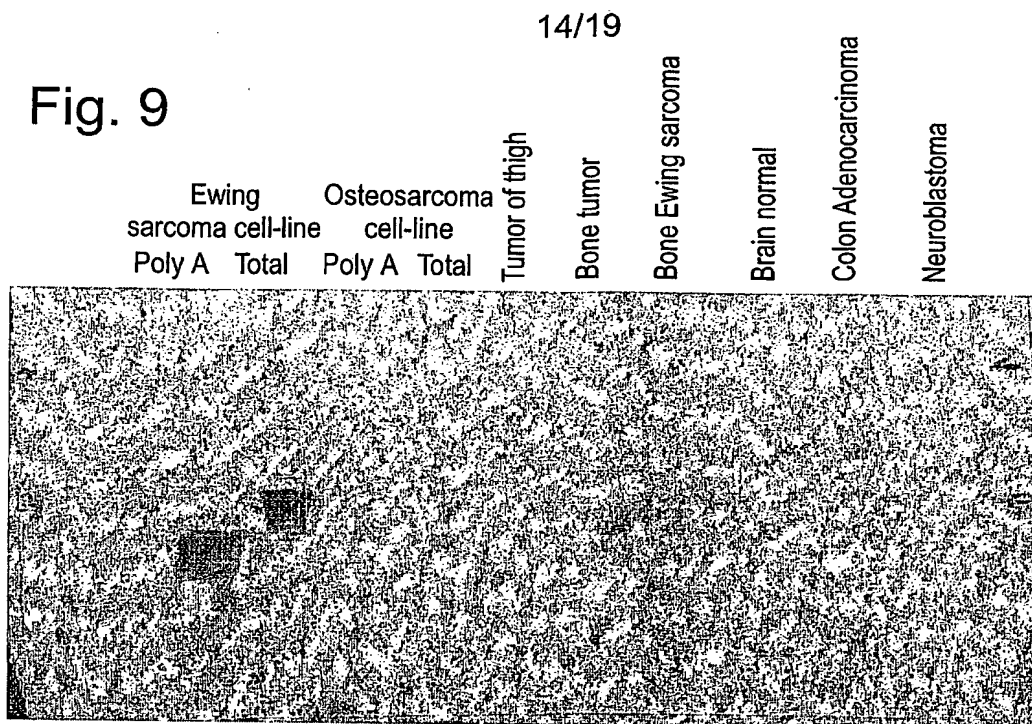


Fig. 8

Fig. 9



Brain  
Stomach  
Placenta  
Kidney  
Spleen  
Thymus  
Adenocarcinoma  
Colon normal  
Colon Duke's A  
Colon Duke's B  
Colon Duke's C  
Colon Duke's D

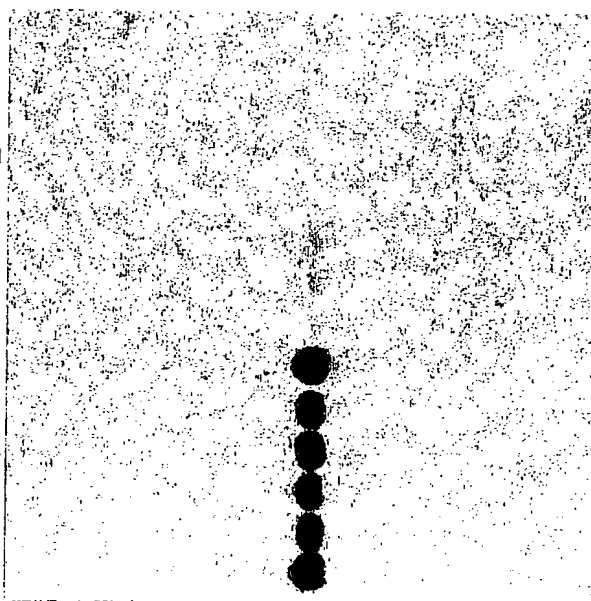


Fig. 10

15/19

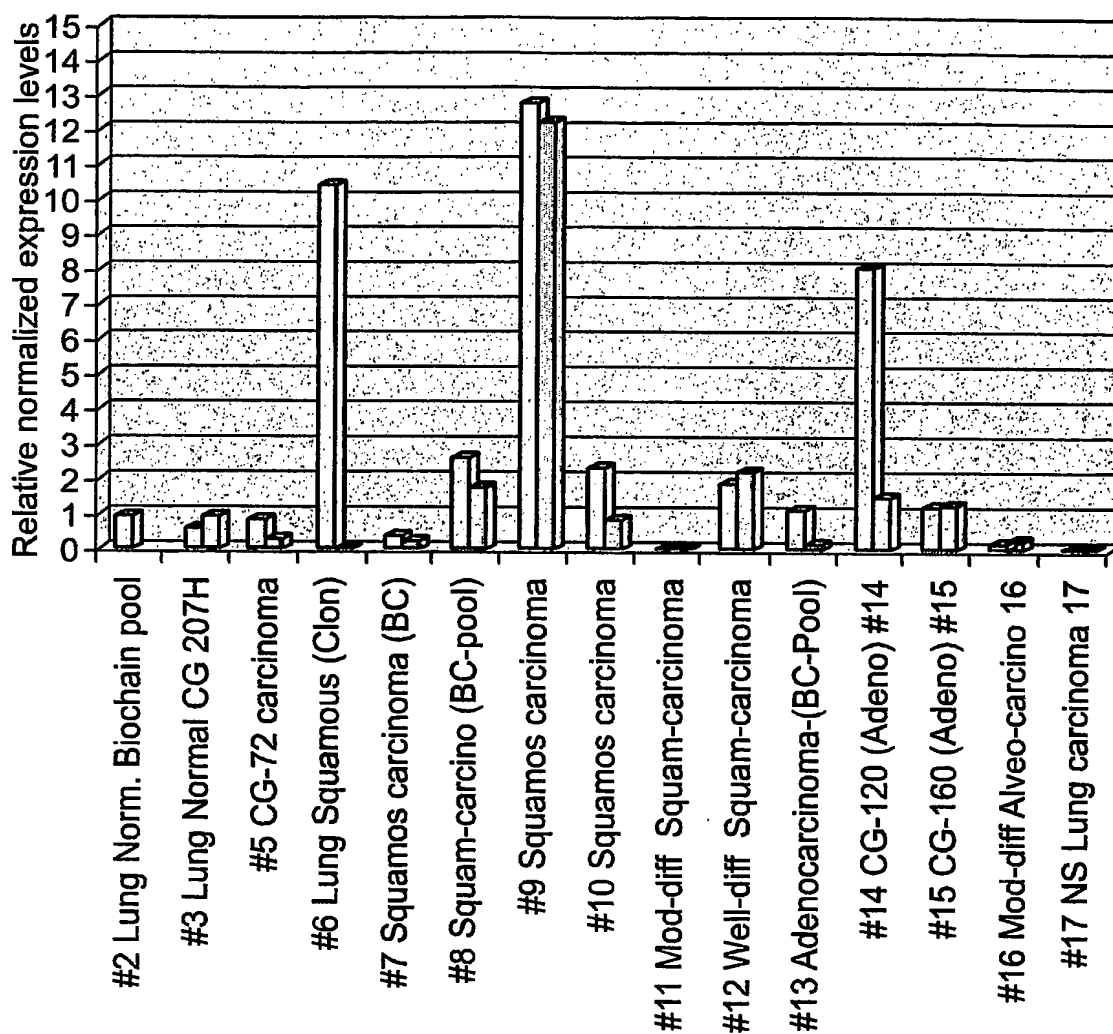


Fig. 11

16/19

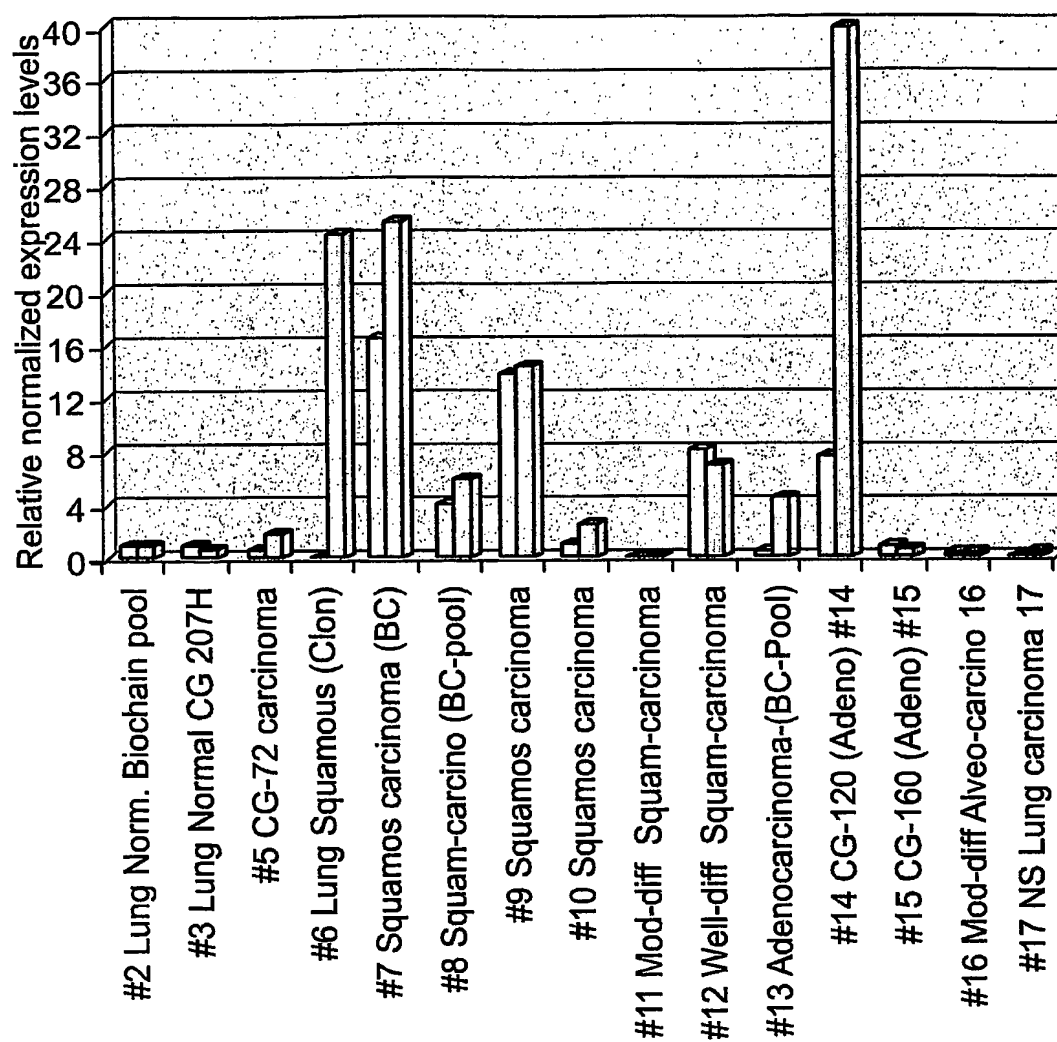


Fig. 12



17/19

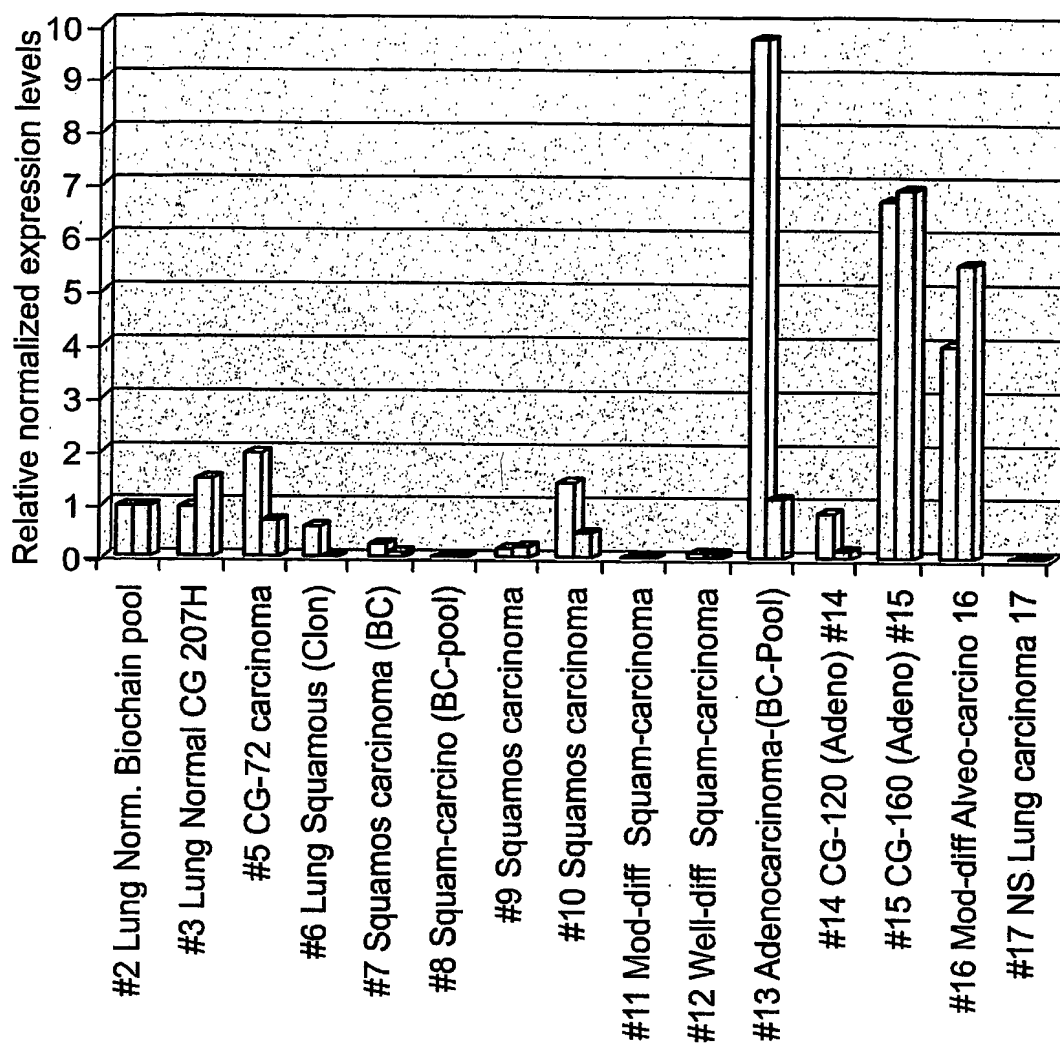
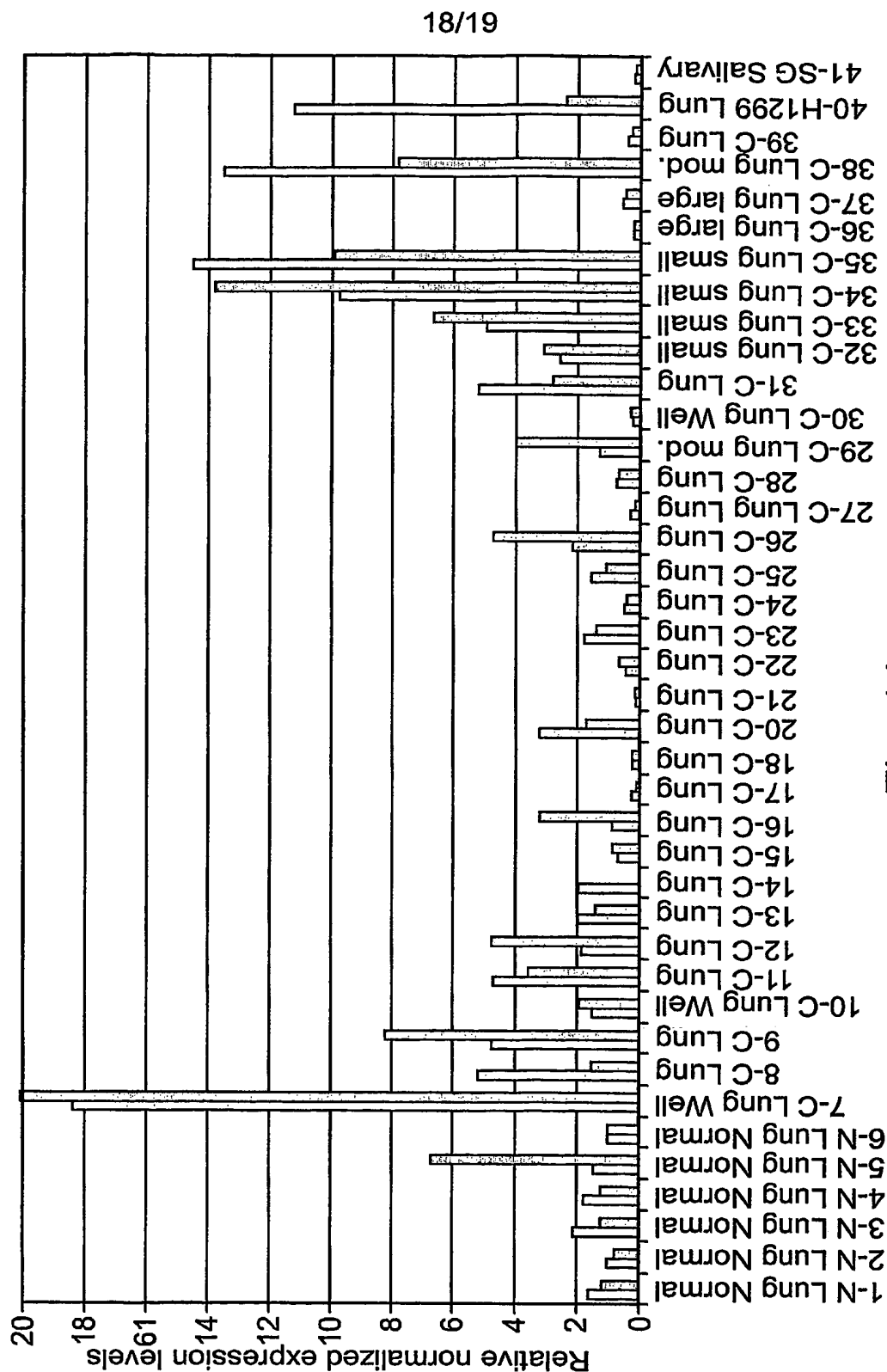
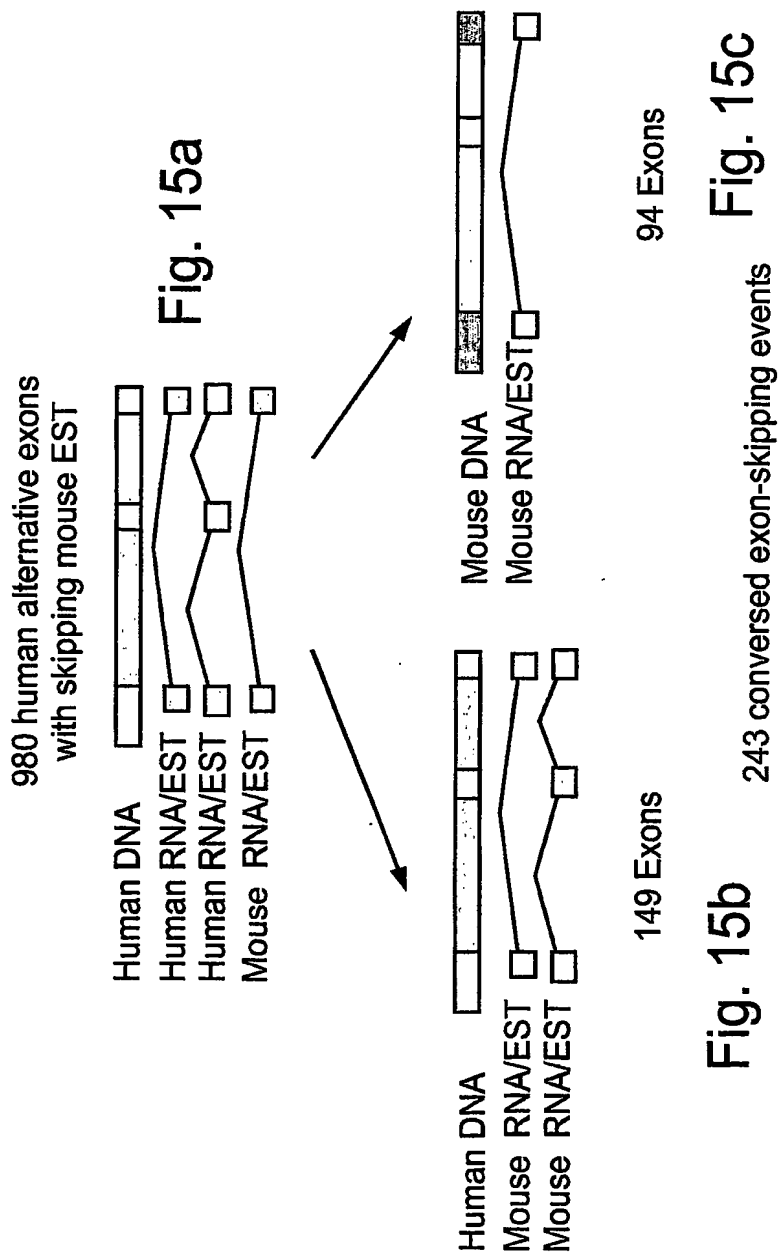


Fig. 13





**This Page is Inserted by IFW Indexing and Scanning  
Operations and is not part of the Official Record**

**BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☒ FADED TEXT OR DRAWING
- ☒ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☒ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☒ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: \_\_\_\_\_

**IMAGES ARE BEST AVAILABLE COPY.**

**As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.**